4-2019

Inference in partially identified panel data models with interactive fixed effects

Shengjie HONG

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

Yaqi WANG

Citation
HONG, Shengjie; SU, Liangjun; and WANG, Yaqi. Inference in partially identified panel data models with interactive fixed effects. (2019). 1-53. Research Collection School Of Economics.
Available at: https://ink.library.smu.edu.sg/soe_research/2286

SMU ECONOMICS & STATISTICS

Inference in Partially Identified Panel Data Models with Interactive Fixed Effects

Shengjie Hong, Liangjun Su, Yaqi Wang

April 2019

Paper No. 14-2019

Inference in Partially Identified Panel Data Models with Interactive Fixed Effects*

Shengjie Hong[a], Liangjun Su[b], Yaqi Wang[c]

[a] School of Economics and Management, Tsinghua University

[b] School of Economics, Singapore Management University

[c] School of Finance, Central University of Finance and Economics

April 19, 2019

Abstract

In this paper we develop methods for statistical inferences in a partially identified nonparametric panel data model with endogeneity and interactive fixed effects. We consider the case where the number of cross-sectional units (N) is large and the number of time series periods (T) as well as the number of unobserved common factors (R) are fixed. Under some normalization rules, we can concentrate out the large dimensional parameter vector of factor loadings and specify a set of conditional moment restrictions that are involved with only the finite dimensional factor parameters along with the infinite dimensional nonparametric component. For a conjectured restriction on the parameter, we consider testing the null hypothesis that the restriction is satisfied by at least one element in the identified set and propose a test statistic based on a novel martingale difference divergence (MDD) measure for the distance between a conditional expectation object and zero. We derive the limiting distribution of the resultant test statistic under the null and show that it is divergent at rate-N under the global alternative based on the U-process theory. To obtain the critical values for our test, we propose a version of multiplier bootstrap and establish its asymptotic validity. Simulations demonstrate the finite sample properties of our inference procedure. We apply our method to study Engel curves for major nondurable expenditures in China by using a panel dataset from the China Family Panel Studies (CFPS).

JEL Classification: C12, C14, C23, C26.

Keywords: Endogeneity, Gaussian chaos process, martingale difference divergence, multiplier bootstrap, nonparametric IV, partial identification, U-processes.

*Su acknowledges the funding support provided by the Lee Kong Chian Fund for Excellence. We thank Andres Santos for very helpful comments and suggestions. Address Correspondence to: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; Phone: +65 6828 0386. ljsu@smu.edu.sg (L. Su).

1 Introduction

Recently there has been a growing interest on panel data models with interactive fixed effects (IFEs). Under a linear specification of the regression relationship, these models have been extensively studied in the literature; see, Coakley, Fuertes and Smith (2002), Phillips and Sul (2003, 2007), Pesaran (2006), Kapetanios and Pesaran (2007), Greenaway-McGrevy, Han and Sul (2008), Bai (2009), Pesaran and Tosetti (2011), Moon and Weidner (2015, 2017), and Lu and Su (2016), among others. More recently, in an effort to relax the linear specification, much attention has been turned to the study of nonparametric panel data models with interactive-effects. See, e.g., Su and Jin (2012), Su, Jin and Zhang (2015), Freyberger (2018), Dong, Gao and Peng (2018), and Su and Zhang (2018) for an overview. In particular, Freyberger (2018) studies a very general nonparametric and nonseparable panel model with IFEs. Nevertheless, all of these papers restrict the covariates to be either strictly or weakly exogenous and assume that the model parameters are point-identified.

In this paper we consider the following nonparametric panel data regression model

$$y_{it} = g^0(x_{it}) + \lambda_i^{0\prime} F_t^0 + u_{it}, \tag{1.1}$$

where $i = 1, ..., N$, $t = 1, ..., T$, x_{it} is a $d_x \times 1$ vector of general regressors with support $\mathcal{X}$, y_{it} is scaler output variable with support $\mathcal{Y}$, F_t^0 and λ_i^0 are $R \times 1$ vectors of unobserved factors and factor loadings, respectively, u_{it} is a zero-mean error term, and the functional form of $g^0(\cdot)$ is unknown. We allow x_{it} and u_{it} to be correlated, and are interested in the inference of $g(\cdot)$ by assuming the presence of a $d_z \times 1$ vector of weakly exogenous instruments z_{it} with support $\mathcal{Z}$, such that

$$\mathbb{E}(u_{it}|z_{i1}, ..., z_{it}) = 0 \text{ almost surely (a.s.).} \tag{1.2}$$

Throughout the paper we assume that T and R are fixed with $T \geq R + 1$, and the asymptotic theory is established by passing N to infinity.

When $g^0(x_{it})$ is linear in x_{it} so that $g^0(x_{it}) = \beta^{0\prime} x_{it}$ for some $\beta^0 \in \mathbb{R}^{d_x}$, x_{it} is strictly exogenous, and $R = 1$, Ahn, Lee and Schmidt (2001) follow the lead of Holz-Eakin, Newey and Rosen (1988) to study the asymptotic properties of the GMM estimator of β based on the quasi-differencing of the equation in (1.1). Ahn, Lee and Schmidt (2013, ALS hereafter) consider the GMM estimation of β when x_{it} is either strictly or weakly exogenous and $R \geq 1$. Su and Jin (2012) study the asymptotic properties of the sieve estimator of g^0 in (1.1) when x_{it} is strictly exogenous. Su and Zhang (2018) consider the sieve estimation of g^0 when x_{it} is weakly exogenous. Freyberger (2018) considers the point-identification and estimation of a model that is more general than that in (1.1), but still restricts x_{it} to be either strictly or weakly exogenous.

The nonparametric IV (NPIV) model for cross-sectional data (i.e., $y_i = g^0(x_i) + u_i$ with $\mathbb{E}(u_i|z_i) = 0$ a.s.) has been widely studied in the literature. NPIV is encompassed by our model as a special case where $T = 1$ and $R = 0$. Point-identification of $g^0(\cdot)$ in NPIV relies heavily on a completeness assumption regarding the joint distribution of x_i and z_i, as formalized by Newey and Powell (2003). Nevertheless, Santos (2012) shows that $g^0(\cdot)$ in the NPIV model is only partially identified in general. In Model (1.1) under our investigation, with the IFEs treated as fixed parameters to be handled, nonparametric point-identification of $g^0(\cdot)$ is even harder to achieve without imposing strong ad hoc assumptions. Therefore, we aim at developing an effective inference method for $g^0(\cdot)$ that is consistent under a potential lack of point-identification in this paper.

For a conjectured restriction on $g(\cdot)$, we develop a consistent procedure for testing the hypothesis that the restriction is satisfied by at least one element in the identified set of $g(\cdot)$. A broad group of restrictions can be tested in this way, making the procedure applicable to various inference tasks, including testing model specification and constructing a confidence set for $g^0(\cdot)$ at any given point. We derive the limiting distribution of our test statistic under the null and show that it is divergent at rate-N under the global alternative based on the U-process theory. To obtain the critical values for our test, we propose a version of multiplier bootstrap and establish its asymptotic validity. We conduct Monte Carlo simulations to demonstrate the finite sample properties of our inference procedure.

Our test statistic is based on a novel martingale difference divergence (MDD) measure for the distance between a conditional expectation object and zero. This way of constructing statistic for testing conditional moment specification is rather different from the widely adopted method, dating back to Bierens (1982), that involves transforming conditional moments into infinitely many unconditional ones via a family of instrument functions and then constructing Kolmogorov Smirnov (KS) or Cramér-von Mises (CvM) type statistic over the instrument function family. As shown in the paper, under partial identification, our MDD-based statistic has two main advantages over Bierens-type statistics: (i) Computing our MDD based statistic is relatively simple regardless of the dimension of the conditioning variables (i.e., $\underline{z}_{it} \equiv (z'_{i1}, ..., z'_{it})'$ for $t = 1, ..., T$ in our model). So it does not suffer from high computational cost even if the dimension of the conditioning variables is moderately large; (ii) The null asymptotic distribution of our statistic is free of any drifting terms, which commonly appear in the null asymptotic distributions of many existing test statistics built to be robust against partial identification (e.g., Santos (2012) and Hong (2017)). This makes it straightforward to implement our testing procedure to obtain an asymptotically exact test.[1]

The main technical challenges of our analysis arise largely because our test statistic is associated with a second order U-process asymptotically that is degenerate under the null and non-degenerate under the alternative. First, out test statistic can be written as a minimizer of a MDD-based process indexed by $\theta = (\phi', g)'$ where ϕ is a finite-dimensional vector associated to the unobserved factor $F = (F_1, ..., F_T)'$ and g is the infinite dimensional parameter of interest. The MDD-based process is a third order U-process that can be decomposed into the summation of a bias term, a second order canonical U-process and a third order canonical U-process via standard Hoeffding decompositions. To study these canonical U-process components, we find helpful insights from de la Peña and Giné (1999) who state some weak convergence results for canonical U-processes with kernel functions belonging to the VC-subgraph class. Such results are not directly applicable to our setting because the kernel functions of our U-processes do not belong to the VC-subgraph class due to the presence of the infinite dimensional parameter $g(\cdot)$. Fortunately, we can verify some primitive conditions in Acrones and Giné (1993) to show that the second order canonical U-process in our Hoeffding decomposition converges weakly to a Gaussian chaos process and the third order term asymptotically vanishes. To the best of our knowledge, such results are the first ones for degenerate U-processes indexed by a non-VC-subgraph class, which complements the literature in both econometrics and statistics.

[1] Santos (2012) employs additional sieve parameters to mimic the corresponding drafting term in his bootstrap procedure so that the resulting test is exact. In comparison, Hong (2017) sets the corresponding drafting term to zero in his bootstrap procedure, which saves computational cost, but leads to potentially conservative tests.

Second, to derive the null asymptotic distribution for our statistic, we borrow ideas from the growing literature on nonparametric partial identification; see, Santos (2012), Andrews and Shi (2014), Chernozhukov, Newey and Santos (2015), and Hong (2017), among others. Our study complements this literature and is closely related to Santos (2012) and Hong (2017). The big difference is that Santos (2012) and Hong (2017) establish their limiting distributions based on standard empirical process theory while we establish our asymptotic results based on the U-process theory. As a first step, we manage to show that any minimizer $\hat{\theta}_N$ of our MDD-based process lies in the $o_p(N^{-1/4})$ neighborhood of the identified set under the L^2-norm. Then we show that our test statistic converges to the minimum of a well-defined (noncentered) Gaussian chaos process without the usual drifting terms that appear in Santos (2012) and Hong (2017). Given the fact that our test statistic is not asymptotically pivotal, we propose a multiplier bootstrap procedure to obtain the bootstrap p-values for asymptotic inference. To show the asymptotic validity of the bootstrap procedure, an essential step is the study of the unconditional central limit theorem (CLT) for the underlying U-process of our bootstrap statistic, which is analogous to the unconditional multiplier CLT for empirical processes studied in van der Vaart and Wellner (1996) and Kororok (2008). It extends the unconditional multiplier CLT for degenerate second order U-statistics in Leucht and Neumann (2013) to degenerate second order U-processes.

As an empirical illustration, we apply our method to study Engel curves for four major nondurable expenditures in China by using a panel dataset from the China Family Panel Studies (CFPS). One of our interesting findings is that, even with a nonparametric specification on $g(\cdot)$, the model does not suffice to adequately describe the Engel curve for food consumption among urban households in China when setting the number R of factors to be 0 or 1. While in comparison our test fails to reject the log-linear specification for the Engel curves among rural households when setting $R = 1$. Our empirical study suggests a difference in the degree of heterogeneity on consumption patterns between within the urban population and within the rural population in China. It also suggests that, even a nonparametric specification on $g(\cdot)$, as general as it is, might still be insufficient to compensate for an inadequate handling of heterogeneity to make the corresponding Engel curve a correctly specified one. These results provide some new insights to the huge literature on empirical studies of Engel curves.

The rest of the paper is organized as follows. In Section 2, we introduce the model, the moment conditions and the hypotheses. In Section 3, we construct the MDD-based test statistic, derive its asymptotic behavior, and propose a consistent multiplier bootstrap procedure to obtain the p-values. In Section 4, we study the finite sample performance of our inference procedure by Monte Carlo simulations. In Section 5, we apply our method to study Chinese households' Engel curves. Final remarks are contained in Section 6. The proofs of all theorems and lemmas are delegated to Appendix A. Additional materials are provided in the online supplementary Appendices B and C.

NOTATION. For a vector or matrix A, we denote its transpose as A' and its Frobenius norm as $|A|$ ($\equiv [\mathrm{tr}(AA')]^{1/2}$), where $\equiv$ means "is defined as". We use $\|\cdot\|$ to denote generic (pseudo) norm. For example, for $\theta = \left(\phi', g\right)'$ where ϕ is a finite-dimensional vector to be specified later on and g is the infinite dimensional parameter, we define $\|\theta\| \equiv |\phi| + \|g\|$ to denote a generic (pseudo) norm for $\theta = \left(\phi', g\right)'$, and one popular choice for $\|\cdot\|$ is the L^2 norm, yielding $\|\theta\|_{L^2} = |\phi| + \|g\|_{L^2}$. The true value of $\theta = \left(\phi', g\right)'$ is denoted as

$\theta^0 = \left(\phi^{0\prime}, g^0\right)'$. The operator $\xrightarrow{p}$, $\xrightarrow{\mathcal{L}}$, and $\Longrightarrow$ denote convergence in probability, weak convergence, and convergence in law in the sense of Chapter 1.3 in van der Vaart and Wellner (1996), respectively.

2 The model and hypotheses

Let $X_i = (x_{i1}, ..., x_{iT})'$, $Y_i = (y_{i1}, ..., y_{iT})'$, $Z_i = (z_{i1}, ..., z_{iT})'$, $\mathbf{g}^0(X_i) = \left(g^0(x_{i1}), ..., g^0(x_{iT})\right)'$, $F^0 = \left(F_1^0, ..., F_T^0\right)'$, and $U_i = (u_{i1}, ..., u_{iT})'$. We can rewrite Model (1.1) with Condition (1.2) in vector form:

$$Y_i = \mathbf{g}^0(X_i) + F^0\lambda_i^0 + U_i \text{ with } \mathbb{E}\left(u_{it}|\underline{z}_{it}\right) = 0 \text{ a.s.,} \tag{2.1}$$

where $\underline{z}_{it} \equiv (z_{i1}', ..., z_{it}')'$.

2.1 The moment condition

To proceed, we show that Model (2.1) is equivalent to a number of conditional moment equations. Like in a typical factor model, F^0 and λ_i^0 are not separately identifiable without restrictions.[2] So as a first step, to rule out such trivial non-identification, we make the normalization assumption that the $T \times R$ matrix F takes a form similar to ALS and Freyberger (2018), as follows:

$$F = \begin{pmatrix} \Phi \\ -I_R \end{pmatrix} \tag{2.2}$$

where Φ is a $(T-R) \times R$ matrix of unrestricted parameters, and (2.2) imposes R^2 restrictions by requiring the last R rows of F to be $-I_R$. Let $\phi = \text{vec}(\Phi') \equiv (\phi_1', ..., \phi_{T-R}')'$, where ϕ_t denotes the tth column of Φ' for $t = 1, ..., T-R$.

Then we define the $T \times (T-R)$ matrix:

$$H(\phi) \equiv \begin{pmatrix} I_{T-R} \\ \Phi' \end{pmatrix} \equiv \left[H_1(\phi_1), ..., H_{T-R}(\phi_{T-R})\right]. \tag{2.3}$$

Note that

$$H\left(\phi^0\right)' F^0 = \left(I_{T-R}, \Phi^0\right)\begin{pmatrix} \Phi^0 \\ -I_R \end{pmatrix} = \mathbf{0}_{(T-R)\times R}, \tag{2.4}$$

where $\phi^0 = \text{vec}\left(\Phi^{0\prime}\right) = (\phi_1^{0\prime}, ..., \phi_{T-R}^{0\prime})'$ denotes the true value of ϕ. Consequently, premultiplying both sides of (2.1) by $H\left(\phi^0\right)'$ helps to eliminate the incidental parameters $\left\{\lambda_i^0\right\}$ from the equation:

$$H\left(\phi^0\right)' Y_i = H\left(\phi^0\right)' \mathbf{g}^0(X_i) + H\left(\phi^0\right)' U_i, \tag{2.5}$$

where

$$H(\phi)' U_i = \begin{pmatrix} H_1(\phi_1)' U_i \\ H_2(\phi_2)' U_i \\ \vdots \\ H_{T-R}(\phi_{T-R})' U_i \end{pmatrix} = \begin{pmatrix} u_{i1} + \phi_1' \dot{U}_i \\ u_{i2} + \phi_2' \dot{U}_i \\ \vdots \\ u_{i,T-R} + \phi_{T-R}' \dot{U}_i \end{pmatrix}$$

[2] This is because $F^0\lambda_i^0 = F^0C^{-1}C\lambda_i^0 = F^*\lambda_i^*$ for any nonsingular matrix C where $F^* = F^0C^{-1}$ and $\lambda_i^* = C\lambda_i^0$. To identify F and $\Lambda = (\lambda_1, ..., \lambda_N)'$, we need to impose R^2 restrictions.

and $\dot{U}_i = (u_{i,T-R+1}, u_{i,T-R+2}, ..., u_{i,T})'$. Let

$$\begin{aligned}
\mathbf{m}(Y_i, \phi, \mathbf{g}(X_i)) &\equiv H(\phi)' [Y_i - \mathbf{g}(X_i)] \\
&= \begin{pmatrix} H_1(\phi_1)' [Y_i - \mathbf{g}(X_i)] \\ H_2(\phi_2)' [Y_i - \mathbf{g}(X_i)] \\ \vdots \\ H_{T-R}(\phi_{T-R})' [Y_i - \mathbf{g}(X_i)] \end{pmatrix} \equiv \begin{pmatrix} m_1(Y_i, \phi_1, \mathbf{g}(X_i)) \\ m_2(Y_i, \phi_2, \mathbf{g}(X_i)) \\ \vdots \\ m_{T-R}(Y_i, \phi_{T-R}, \mathbf{g}(X_i)) \end{pmatrix}. \qquad (2.6)
\end{aligned}$$

Then under the condition that the instrument z_{it} is weakly exogenous, we can easily see that

$$\mathbb{E}\left[m_s(Y_i, \phi_s^0, \mathbf{g}^0(X_i)) | \underline{z}_{is}\right] = 0 \text{ a.s. for } s = 1, ..., T-R. \qquad (2.7)$$

When ϕ^0 and g^0 are point-identified, various methods have been proposed to study the estimation of ϕ and g in the above model. See Ai and Chen (2003) and Chen and Pouzo (2012), among others.

2.2 The parameter space Θ

The parameter space Θ for $\theta = (\phi', g)'$ is specified as $\Theta = \mathbf{\Phi} \times \mathcal{G}$ as in Hong (2017), where $\mathbf{\Phi}$ is a compact subset of $\mathbb{R}^{(T-R)R}$ and $\mathcal{G}$ is a bounded subset of the following Sobolev space:

$$\mathcal{W}^s(\mathcal{X}) \equiv \{g : \mathcal{X} \to \mathbb{R} \mid \ g \text{ is } d\text{-times differentiable and } \|g\|_s \leq \infty\}$$

with $\|\cdot\|_s$ being a commonly used norm for weighted Sobolev spaces, defined as

$$\|g\|_s^2 \equiv \sum_{\langle\lambda\rangle \leq d} \int_{\mathcal{X}} \left|D^\lambda g(x)\right|^2 (1 + x'x)^{\zeta_0} \, dx$$

where $\lambda \in \mathbb{N}_+^{d_x}$, $\langle\lambda\rangle \equiv \sum_{j=1}^{d_x} \lambda_j$, $D^\lambda g(x) \equiv \partial^{\langle\lambda\rangle} g(x) / \prod_{j=1}^{d_x} \partial x_j^{\lambda_j}$, $d \in \mathbb{N}_+$ measures the degree of smoothness, and $\zeta_0 \geq 0$. Define another norm $\|\cdot\|_c$ as follows

$$\|g\|_c \equiv \max_{\langle\lambda\rangle \leq \frac{d}{2}} \left[\sup_{x \in \mathcal{X}} \left|D^\lambda g(x)\right| (1 + x'x)^{\zeta/2}\right]$$

with $\zeta = 0$ for bounded $\mathcal{X}$, and $\left(\frac{d_x}{2} \lfloor \frac{d}{2} \rfloor\right) / \left(\lfloor \frac{d}{2} \rfloor - \frac{d_x}{2}\right) < \zeta < \zeta_0$ for unbounded $\mathcal{X}$. Here, $\lfloor a \rfloor$ represents the largest integer that is no larger than a.

To be precise, we specify the parameter space $\Theta = \mathbf{\Phi} \times \mathcal{G}$ as follows:

Assumption 2.1 *(i) $\mathbf{\Phi} \subset \mathbb{R}^{(T-R)R}$ is compact; (ii) $\mathcal{G} = \{g \in \mathcal{W}^s(\mathcal{X}) : \|g\|_s \leq B\}$ for some $B < \infty$ and $d \geq d_x + 2$. $\zeta_0 = 0$ for bounded $\mathcal{X}$, and $\zeta_0 > \left(\frac{d_x}{2} \cdot \lfloor \frac{d}{2} \rfloor\right) / \left(\lfloor \frac{d}{2} \rfloor - \frac{d_x}{2}\right)$ for unbounded $\mathcal{X}$; (iii) let*

$$\nu \equiv \begin{cases} (d - \lfloor d/2 \rfloor + \zeta) \, d_x / \{\zeta(d - \lfloor d/2 \rfloor)\} & \text{if } \mathcal{X} \text{ is unbounded,} \\ d_x / (d - \lfloor d/2 \rfloor) & \text{if } \mathcal{X} \text{ is bounded.} \end{cases} \qquad (2.8)$$

Then $\nu < 1$; (iv) $\mathcal{X}$ satisfies a uniform cone condition.

Remark. Assumption 2.1(i) is standard and Assumption 2.1(ii)-(iii) parallels Assumption 2.1(i)-(ii) in Santos (2012). Assumption 2.1(ii) specifies $\mathcal{G}$ to be a bounded ball under $\|\cdot\|_s$ with radius B in $\mathcal{W}^s(\mathcal{X})$. Such a specification brings the following benefits: (i) As noted by Santos (2012), $\mathcal{G}$ is compact under the norm $\|\cdot\|_c$. Consequently, $\Theta = \mathbf{\Phi} \times \mathcal{G}$ is compact under $\|\cdot\|_c$ defined on $\mathbb{R}^{(T-R)R} \times \mathcal{W}^s(\mathcal{X})$ as

$$\|\theta\|_c \equiv |\phi| + \|g\|_c \tag{2.9}$$

for $\theta = (\phi', g)'$. (We formalize this compactness result for Θ as Lemma A.3 in the Appendix.) It immediately follows that Θ is compact under any norm that is weaker than $\|\cdot\|_c$, such as $\|\cdot\|_s$ and $\|\cdot\|_{L^2}$; (ii) As noted by Hong (2017), the compactness of Θ under $\|\cdot\|_s$ makes the results in Schumaker (2007) applicable to developing primitive conditions for the required uniform rate (over Θ) of sieve approximation errors (to be specified by Assumption 3.3 later in the paper).

2.3 Hypotheses and notion of test

Define

$$\Theta_I \equiv \{\theta = (\phi', g)' \in \mathbf{\Phi} \times \mathcal{G} : \mathbb{E}\left[m_s(Y_i, \phi_s, \mathbf{g}(X_i)) | \underline{z}_{is}\right] = 0 \text{ a.s. for } s = 1, ..., T-R\}. \tag{2.10}$$

Θ_I is referred to as the identified set in the literature. We say that $\theta = (\phi', g)'$ is partially identified by (2.7) if Θ_I contains more than one element. The following lemma suggests that there is no loss of information by considering Θ_I defined in (2.10) instead of (1.1)–(1.2), i.e., the original model.

Lemma 2.1 (No loss of information) *Θ_I, the identified set defined by (2.10), is the same as the identified set characterized by (1.1)–(1.2). That is, Θ_I is equivalent to*

$$\left\{ \theta = (\phi', g)' \in \Theta : \begin{array}{l} \text{For some } R\text{-dimensional random vector } \lambda_i, \text{ it holds} \\ \mathbb{E}\left[y_{it} - g(x_{it}) - \lambda_i'\phi_t | \underline{z}_{it}\right] = 0 \text{ a.s. for } t = 1, ..., T-R \\ \mathbb{E}\left[y_{it} - g(x_{it}) - \lambda_i'\left(-\iota_{t-(T-R)}\right) | \underline{z}_{it}\right] = 0 \text{ a.s. for } t = T-R+1, ..., T \end{array} \right\}$$

where ι_t represents the t'th column of the $R \times R$ identity matrix.

For hypothesis testing on a conjectured restriction on θ, in the generic form

$$L(\theta) = l,$$

we consider testing whether such a restriction is satisfied by at least one element of the identified set. Equivalently, defining the restricted set as $\Theta_R \equiv \{\theta \in \Theta : L(\theta) = l\}$, the null and alternative hypotheses under our consideration are

$$\mathbb{H}_0 : \Theta_I \cap \Theta_R \neq \emptyset \quad \text{v.s.} \quad \mathbb{H}_1 : \Theta_I \cap \Theta_R = \emptyset,$$

where $\emptyset$ denotes the empty set.

The notion of the above testing hypotheses is widely adopted under partial identification. When $\theta^0 = (\phi^{0\prime}, g^0)'$ is point-identified by (2.7), the above null hypothesis $\mathbb{H}_0$ simply tests whether θ^0 satisfies the specified restriction in $\Theta_R : L(\theta^0) = l$.

We consider the same family of restrictions as in Santos (2012) and Hong (2017):

Assumption 2.2 *For $(\mathcal{L}, \|\cdot\|_{\mathcal{L}})$ a Banach space, $L : (\mathcal{G}, \|\cdot\|_c) \to (\mathcal{L}, \|\cdot\|_{\mathcal{L}})$ is a bounded linear operator.*

As discussed in Santos (2012), Assumption 2.2 appears restrictive but actually encompasses a broad group of restrictions since one can flexibly choose the Banach space $(\mathcal{L}, \|\cdot\|_{\mathcal{L}})$. For example, we can test whether the value of the function g^0 at a point x^0 is given by a value γ^0 by setting $(\mathcal{L}, \|\cdot\|_{\mathcal{L}}) = (\mathbb{R}, |\cdot|)$, $\Theta_R = \{g \in \mathcal{W}^s(\mathcal{X}) : g(x^0) = \gamma^0\}$, $L(g) = g(x^0)$, and $l = \gamma^0$. For another example, we can test whether g^0 is an affine function by setting $(\mathcal{L}, \|\cdot\|_{\mathcal{L}}) = (L^2(X), \|\cdot\|_{L^2})$, $\Theta_R = \{g \in \mathcal{W}^c(\mathcal{X}) : g(x) = \beta_0 + \beta_1' x$ for some $(\beta_0, \beta_1')' \in \mathbb{R}^{d_x+1}\}$, $L(g) = g - P_{\mathcal{A}}(g)$, and $l = 0$. Here, $L^2(X) = \{b : \mathcal{X} \to \mathbb{R} : \mathbb{E}[b(X)^2] < \infty\}$, $\|b\|_{L^2} = \{\mathbb{E}[b(X)^2]\}^{1/2}$, and $P_{\mathcal{A}}(g)$ denote the projection of $g \in L^2(X)$ onto $\mathcal{A} \equiv \{g \in W^c(\mathcal{X}) : g(x) = \beta_0 + \beta_1' x$ for some $(\beta_0, \beta_1')' \in \mathbb{R}^{d_x+1}\}$. For additional examples of restrictions that satisfy Assumption 2.2, see Santos (2012).

In the panel data model with IFEs, $g(\cdot)$, or its functional, is typically the parameter of interest, in which case ϕ can be regarded as a nuisance parameter. For this reason, we mainly consider hypotheses that impose restrictions on $g(\cdot)$ alone, in which case the restriction to be tested takes the special form $L(g) = l$. Then the restricted set becomes $\Theta_R = \{\theta = (\phi, g) \in \mathbf{\Phi} \times \mathcal{G} : L(g) = l\}$.

3 The testing procedure

3.1 Test statistics

A popular method to handle hypothesis testing for conditional moment models is to construct test statistics based on equivalent unconditional moments. This method dates back to Bierens (1982) and has been adopted in many papers on point-identification analysis (see, e.g., Stinchcombe and White (1998) and Dominguez and Lobato (2004)), and in more recent papers on partial identification analysis such as Santos (2012), Andrews and Shi (2013), and Hong (2017). To adopt this method in our study, it requires the choice of a family of generically revealing functions $(\varphi_1(t_1, \cdot), \varphi_2(t_2, \cdot), \ldots \varphi_{T-R}(t_{T-R}, \cdot))$ indexed by $\mathbf{t} \equiv (t_1', t_2', \ldots, t_{T-R}')' \in \prod_{s=1}^{T-R} \mathcal{T}_s \equiv \mathcal{T}$ that satisfies the following condition

$$\mathbb{E}\left[m_s(Y_i, \phi_s, \mathbf{g}(X_i)) | \underline{z}_{is}\right] = 0 \text{ a.s. iff } \mathbb{E}\left[m_s(Y_i, \phi_s, \mathbf{g}(X_i)) \varphi_s(t_s, \underline{z}_{is})\right] = 0 \text{ for all } t_s \in \mathcal{T}_s.$$

Then we can construct the following test statistic

$$\bar{J}_N \equiv \min_{\theta \in \Theta_N \cap \Theta_R} \max_{\mathbf{t} \in \mathcal{T}} N \cdot |J_N(\theta, \mathbf{t})|^2, \tag{3.1}$$

where

$$J_N(\theta, \mathbf{t}) \equiv \begin{pmatrix} \frac{1}{N}\sum_{i=1}^N m_{i1}(\phi_1, g)\, \varphi_1(t_1, \underline{z}_{i1}) \\ \frac{1}{N}\sum_{i=1}^N m_{i2}(\phi_2, g)\, \varphi_2(t_2, \underline{z}_{i2}) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^N m_{i,T-R}(\phi_{T-R}, g)\, \varphi_{T-R}(t_{T-R}, \underline{z}_{i,T-R}) \end{pmatrix} \equiv \begin{pmatrix} J_{N1}(\phi_1, g, t_1) \\ J_{N2}(\phi_2, g, t_2) \\ \vdots \\ J_{N,T-R}(\phi_{T-R}, g, t_{T-R}) \end{pmatrix},$$

with $m_{is}(\phi_s, g) \equiv m_s(Y_i, \phi_s, \mathbf{g}(X_i))$ and Θ_N is an approximating sieve space for Θ.

Hong (2017) shows under a general setting that statistics of the form (3.1) weakly converge to certain functional of a Gaussian process under the null, and proposes a penalized bootstrap procedure for testing.

These results are potentially applicable to the statistic $\bar{J}_N$ in our study. However, note that the dimension of $\underline{z}_{is}$ is given by sd_z, and the dimension of the index $\mathbf{t}$ is typically equal or comparable to $\left(\sum_{s=1}^{T-R} s\right) d_z$, which can get relatively large for moderate sizes of $T-R$ and d_z. As a result, the computation of $\bar{J}_N$ can be extremely expansive. For this reason, we opt for a different test statistic based on the notion of MDD.

An MDD-based statistic for our study is motivated by two recent papers: Shao and Zhang (2014) and Su and Zheng (2017). For any real-valued variable V and vector-valued variable W, the original version of MDD, denoted by $\mathrm{MDD}_o\,(V|W)^2$, is defined by Shao and Zhang (2014) as

$$\mathrm{MDD}_o(V|W)^2 \equiv -\mathbb{E}\left\{[V-\mathbb{E}(V)]\left[V^\dagger-\mathbb{E}\left(V^\dagger\right)\right]\left|W-W^\dagger\right|\right\}, \tag{3.2}$$

where $(V^\dagger, W^\dagger)$ is an independent copy of (V, W). Shao and Zhang (2014) show that under some suitable moment conditions ($\mathbb{E}(V^2)<\infty$ and $0<\mathbb{E}(|W|^2)<\infty$), $\mathrm{MDD}_o\,(V|W)^2 \geq 0$ and

$$\mathrm{MDD}_o\,(V|W)^2 = 0 \text{ iff } \mathbb{E}(V|W)=\mathbb{E}\,(V) \quad \text{a.s.} \tag{3.3}$$

Based on the above properties, they propose a consistent test for conditional mean independence condition of (3.3).

Su and Zheng (2017) propose a modified version of MDD, defined as:

$$\mathrm{MDD}\,(\varepsilon|W)^2 \equiv -\mathbb{E}\left[\varepsilon\varepsilon^\dagger\left|W-W^\dagger\right|\right]+2\mathbb{E}\left[\varepsilon\left|W-W^\dagger\right|\right]\mathbb{E}\left[\varepsilon^\dagger\right], \tag{3.4}$$

where ε is the error term of a nonlinear regression model, W is the regressor, and $\left(\varepsilon^\dagger, W^\dagger\right)$ is an independent copy of (ε, W). Su and Zheng (2017) show that $\mathrm{MDD}\,(\varepsilon|W)^2 \geq 0$ and

$$\mathrm{MDD}\,(\varepsilon|W)^2 = 0 \text{ iff } \mathbb{E}(\varepsilon|W)=0 \quad \text{a.s.} \tag{3.5}$$

Then they propose a novel and effective test for correct (parametric) specification of the regression function based on the above properties of MDD. Notably, in their setting, the regression function is correctly specified if and only if the conditional mean zero condition of (3.5) holds. Simulation results in Su and Zheng (2017) indicate that a test statistic based on the MDD significantly outperforms many popular specification tests in the literature, and that it performs well even when the dimension of W is large.

Following the insights from Su and Zheng (2017), it can be shown, under our setting of Θ and Θ_R, that $\Theta_I \cap \Theta_R \neq \emptyset$ if and only if:

$$\min_{\theta\in\Theta\cap\Theta_R}\left\{\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,\phi_s,\mathbf{g}\left(X\right)\right)|\underline{z}_s\right]^2\right\}=0. \tag{3.6}$$

And the construction of the test statistic for

$$\mathbb{H}_0:\Theta_I\cap\Theta_R\neq\emptyset \quad \text{v.s.} \quad \mathbb{H}_1:\Theta_I\cap\Theta_R=\emptyset \tag{3.7}$$

proceeds in two steps as follows:

I. Fix $\theta\in\Theta$ and derive a test statistic $S_N(\theta)$ for the null hypothesis $\mathbb{H}_0:\mathbb{E}\left[m_s\left(Y,\phi_s,\mathbf{g}\left(X\right)\right)|\underline{z}_s\right]=0$ a.s. for all $s=1,...,T-R$, or equivalently, $\mathbb{H}_0:\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,\phi_s,\mathbf{g}\left(X\right)\right)|\underline{z}_s\right]^2=0$ for all $s=1,...,T-R$.

II. Let Θ_N be a sieve approximating space for Θ. Then, following (3.6), test $\mathbb{H}_0 : \Theta_I \cap \Theta_R \neq \emptyset$ by using the statistic $\hat{S}_N = \min_{\theta \in \Theta_N \cap \Theta_R} S_N(\theta)$.

For Step I, we propose the following statistic:

$$S_N(\theta) \equiv \sum_{s=1}^{T-R} S_{Ns}(\theta) \tag{3.8}$$

where $S_{Ns}(\theta)$ is constructed in a way similar to Su and Zheng (2017):

$$S_{Ns}(\theta) = -\frac{1}{N} \sum_{1 \leq i \neq j \leq N} m_{is}(\theta) m_{js}(\theta) \kappa_{ij,s} + \frac{2}{N} \sum_{1 \leq i \neq j \leq N} m_{is}(\theta) \kappa_{ij,s} \frac{1}{N} \sum_{k=1}^{N} m_{ks}(\theta) \tag{3.9}$$

with $m_{is}(\theta) \equiv m_s(Y_i, \phi_s, \mathbf{g}(X_i))$ and $\kappa_{ij,s} = \left| \underline{z}_{is} - \underline{z}_{js} \right|$ for $s = 1, ..., T-R$.

For Step II, we define the test statistic accordingly as

$$\hat{S}_N = \min_{\theta \in \Theta_N \cap \Theta_R} S_N(\theta), \tag{3.10}$$

where $\Theta_N = \mathbf{\Phi} \times \mathcal{G}_N$ is an approximating space for $\Theta = \mathbf{\Phi} \times \mathcal{G}$, with

$$\mathcal{G}_N = \left\{ g \in \mathcal{G} : g(\cdot) = p^{k_N}(\cdot)' \beta \text{ for some } \beta \in \mathbb{R}^{k_N} \right\}$$

being an approximating space for $\mathcal{G}$ by using a k_N-vector of basis functions $p^{k_N}(\cdot) = (p_1(\cdot), ..., p_{k_N}(\cdot))'$ defined on $\mathcal{X}$.

3.2 Definitions and notations

For $\theta^0 = \left(\phi^{0\prime}, g^0\right)' \in \Theta_I \cap \Theta_R$, let

$$\Pi_N \theta^0 \equiv \begin{pmatrix} \phi^0 \\ \Pi_{\mathcal{G}_N} g^0 \end{pmatrix} \tag{3.11}$$

be the projection of θ^0 onto $\Theta_N \cap \Theta_R$.

Definition 3.1 (Weak pseudo-metric) *Let $m_s(Y, X, \theta) \equiv m_s(Y, \phi_s, \mathbf{g}(X))$. Define the following pseudo-metric $\mathrm{d}_w(\cdot, \cdot)$ on Θ:*

$$\mathrm{d}_w(\theta_1, \theta_2) \equiv \left\{ \sum_{s=1}^{T-R} \mathrm{MDD} \left[(m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2)) | \underline{z}_s \right]^2 \right\}^{1/2},$$

for any θ_1, $\theta_2 \in \Theta$.

Note that Θ_I forms an equivalent class under $\mathrm{d}_w(\cdot, \cdot)$, i.e., for any $\theta_1^0 \in \Theta_I$ and $\theta_2^0 \in \Theta_I$, $\mathrm{d}_w\left(\theta_1^0, \theta_2^0\right) = 0$, which is made clear by Lemma A.1 in the Appendix. It also follows from Lemma A.1 that for any given $\theta \in \Theta$ and $\theta^0 \in \Theta_I$,

$$\mathrm{d}_w\left(\theta, \theta^0\right) = \left\{ \sum_{s=1}^{T-R} \mathrm{MDD} \left[m_s(Y, X, \theta) | \underline{z}_s \right]^2 \right\}^{1/2}.$$

The following lemma shows that $\mathrm{d}_w(\cdot, \cdot)$ is weaker than the L^2-metric and satisfies a triangle-like inequality around the identified set.

Lemma 3.1 *Let Assumptions 2.1 hold. Suppose that* $\mathbb{E}[|Y|^2] < \infty$ *and* $\mathbb{E}|Z| < \infty$*. Then there exists a finite constant* $c > 0$ *s.t.* $\mathrm{d}_w(\theta_1, \theta_2) \leq c \|\theta_1 - \theta_2\|_{L^2}$ *for any* θ_1*,* $\theta_2 \in \Theta$*. In addition,* $\mathrm{d}_w(\theta_1, \theta_2) \leq 2\left[\mathrm{d}_w\left(\theta_1, \theta^0\right) + \mathrm{d}_w\left(\theta_2, \theta^0\right)\right]$ *for any* $\theta^0 \in \Theta_I$*.*

Definition 3.2 (Pathwise derivatives) *For the functions* $m_s(Y, X, \cdot) : \Theta \to \mathbb{R}$*,* $s = 1, ..., T - R$*, we define the first-order pathwise derivative of* $m_s(Y, X, \cdot)$ *at* θ^0 *in the direction of* Δ *as*

$$\frac{\partial m_s\left(Y, X, \theta^0\right)}{\partial \theta}[\Delta] = \frac{\partial m_s\left(Y, X, \theta^0 + \tau\Delta\right)}{\partial \tau}|_{\tau=0},$$

and the second-order pathwise derivative of $m_s(Y, X, \cdot)$ *at* θ^0 *in the direction of* Δ *as*

$$\frac{\partial^2 m_s\left(Y, X, \theta^0\right)}{\partial \theta^2}[\Delta, \Delta] = \frac{\partial^2 m_s\left(Y, X, \theta^0 + \tau\Delta\right)}{\partial \tau^2}|_{\tau=0}.$$

Given the functional form of m_s as specified in (2.6), it holds that

$$m_s(Y, X, \theta) - m_s\left(Y, X, \theta^0\right) = \frac{\partial m_s\left(Y, X, \theta^0\right)}{\partial \theta}\left[\theta - \theta^0\right] + \frac{1}{2}\frac{\partial^2 m_s\left(Y, X, \theta^0\right)}{\partial \theta^2}\left[\theta - \theta^0, \theta - \theta^0\right]. \tag{3.12}$$

As formalized in Lemma A.6(i) in Appendix A, under mild conditions (Assumptions 2.1, 2.2, and 3.1 – 3.3(i) to be specified in the next subsection), any minimizer $\hat{\theta}_N$ of $S_N(\theta)$ over Θ_N would lie in the $o_p(1)$ neighborhood of the identified set Θ_I under the L^2-norm. Given this consistency result, we can restrict our attention on a shrinking L^2 sieve neighborhood around Θ_I, defined as

$$\Theta_{oN} \equiv \left\{\theta \in \Theta_N : \mathrm{d}_{\|\cdot\|_{L^2}}(\theta, \Theta_I) \leq \varsigma_N\right\}$$

for some positive $\varsigma_N \downarrow 0$.[3] And we define the following measure of local ill-posedness.

Definition 3.3 (Sieve measure of local ill-posedness) *Define the following sieve measure of local ill-posedness*

$$\varrho_N \equiv \sup_{\theta \in \Theta_{oN} : \theta \notin \Pi_N \Theta_I} \frac{\mathrm{d}_{\|\cdot\|_{L^2}}(\theta, \Pi_N \Theta_I)}{\mathrm{d}_w(\theta, \Pi_N \Theta_I)}.$$

In our analysis, we allow for moderate ill-posedness in the sense that $\varrho_N \uparrow \infty$ but at a slow rate. ϱ_N provides a link between $d_w(\cdot, \cdot)$ and the L^2 distance. So once we establish the rate of convergence under $d_w(\cdot, \cdot)$, a certain rate of convergence under the L^2 distance can be established via ϱ_N. Note that ϱ_N is defined in almost the same way as in Hong (2017), except that our pseudo metric is different from Hong's (2007). Under point-identification, $\varrho_N = \sup_{\theta \in \Theta_{oN} : \theta \neq \Pi_N \theta^0} \frac{\left\|\theta - \Pi_N \theta^0\right\|_{L^2}}{\mathrm{d}_w(\theta, \Pi_N \theta^0)}$, which is similar to the sieve measure of local ill-posedness in Chen and Pouzo (2012).

[3] $\mathrm{d}_{\|\cdot\|_{L^2}}(\theta, \Theta_I) \equiv \inf_{\tilde{\theta} \in \Theta_I} \left\|\theta - \tilde{\theta}\right\|_{L^2}$ represents the distance between a point θ and the set Θ_I under $\|\cdot\|_{L^2}$. Similarly, in Definition 3.3, $\mathrm{d}_w(\theta, \Pi_N \Theta_I) \equiv \inf_{\tilde{\theta} \in \Pi_N \Theta_I} \mathrm{d}_w(\theta, \tilde{\theta})$ represents the distance between a point θ and the set $\Pi_N \Theta_I$ under $\mathrm{d}_w(\cdot, \cdot)$.

3.3 Asymptotic theory

In this section we study the asymptotic properties of $S_N(\theta)$ and $\hat{S}_N$, defined in (3.8) and (3.10), respectively. To establish the asymptotic behavior of $S_N(\theta)$, we impose the following assumption.

Assumption 3.1 *(i) $\{X_i, Y_i, Z_i\}_{i=1}^N$ are i.i.d. with support $\mathcal{X}^T \times \mathcal{Y}^T \times \mathcal{Z}^T$ such that all marginal and joint density functions of X_i, Y_i and Z_i are bounded.*

(ii) $\mathbb{E}[(|Y_i|^2 + 1)(|Z_i|^2 + 1)] < \infty$.

Assumption 3.1(i) is commonly imposed for panel data analyses with individual fixed effects or IFEs. And it does not rule out dynamic panels as along as we treat the unobserved factors F_t^0's as nonrandom. In the case where F_t^0's are random, the independence assumption can be replaced by conditional independence: the lagged dependent variables (e.g., $Y_{i,t-1}$) can be independent across i given the minimal sigma-field generated by the common factors. Assumption 3.1(ii) specifies some moment conditions on Y_i and Z_i.

Let $\xi_i = (Y_i', X_i', Z_i')'$ and $\tilde{m}_{is}(\theta) = m_s(Y_i, X_i, \theta) - \mathbb{E}[m_s(Y_i, X_i, \theta) | \underline{z}_{is}]$. Define the second order U-process indexed by θ as follows:

$$\mathbb{U}_{Ns}(\theta) = \binom{N}{2}^{-1} \sum_{1 \le i < j \le N} h_s(\xi_i, \xi_j; \theta),$$

where $h_s(\xi_i, \xi_j; \theta) = \tilde{m}_{is}(\theta) \tilde{m}_{js}(\theta) [\mathbb{E}_j(\kappa_{ij,s}) + \mathbb{E}_i(\kappa_{ij,s}) - \kappa_{ij,s}]$, and $\mathbb{E}_j(\kappa_{ij,s})$ denotes the expectation with respect to (w.r.t.) the variable $\underline{z}_{js}$ alone in $\kappa_{ij,s} = |\underline{z}_{is} - \underline{z}_{js}|$. Denote by $< \cdot, \cdot >$ the usual inner product on $L^2(P)$, for P a generic probability measure (or a generic marginal one) on $\mathcal{X}^T \times \mathcal{Y}^T \times \mathcal{Z}^T$. For $f \in L^2(P^2) \equiv L^2(P \otimes P)$, define a Hilbert-Schmidt operator H_f on $L^2(P)$ by $(H_f g)(\xi) = Pf(\xi, \cdot) g(\cdot)$. Also, define a process $\mathbb{C}$ on $\mathcal{F}$ by

$$\mathbb{C}(f) = \sum_{\alpha=1}^{\infty} < H_f w_\alpha, w_\alpha > (W_\alpha^2 - 1),$$

where $\{w_\alpha\}$ denotes the eigenfunctions of the operator H_f, and $\{W_\alpha\}$ is a sequence of independent $N(0,1)$ random variables.

The following theorem studies the asymptotic properties of the process $\{S_N(\theta)\}$.

Theorem 3.1 *Let Assumptions 2.1, 2.2, and 3.1 hold. Then*

(i) For each $s = 1, ..., T-R$, $S_{Ns}(\theta) = 2\mathbb{E}[m_s^2(Y, \phi_s, \mathbf{g}(X)) |\underline{z}_s - \underline{z}_s^\dagger|] + \mathbb{U}_{Ns}(\theta) + O_p(N^{-1/2}) \Longrightarrow \mathbb{B}_s(\theta) + \mathbb{C}_s(\theta)$ in $L^\infty(\Theta_I)$ where $\mathbb{B}_s(\theta) = 2\mathbb{E}[m_s^2(Y, \phi_s, \mathbf{g}(X)) |\underline{z}_s - \underline{z}_s^\dagger|]$ and $\mathbb{C}_s(\theta) = \mathbb{C}(h_s(\cdot, \cdot; \theta))$ is a Gaussian chaos process on $L^\infty(\Theta_I)$.[4] $S_N(\theta) \Longrightarrow \sum_{s=1}^{T-R} [\mathbb{B}_s(\theta) + \mathbb{C}_s(\theta)]$ on $L^\infty(\Theta_I)$.

(ii) $\frac{1}{N} S_N(\theta) = \sum_{s=1}^{T-R} \mathrm{MDD}[m_s(Y, X, \theta) | \underline{z}_s]^2 + O_p(N^{-1/2})$ uniformly in $\theta \in \Theta \backslash \Theta_I$.

Theorem 3.1(i) indicates that $S_{Ns}(\theta)$, after being recentered around $\mathbb{B}_s(\theta)$, is essentially a degenerate second order U-process on Θ_I that converges weakly to a Gaussian chaos process $\{\mathbb{C}_s(\theta)\}$. Theorem 3.1(ii) indicates that for $\theta \in \Theta \backslash \Theta_I$, $S_N(\theta)$ is dominated by its deterministic component that is associated to the MDD measure.

To establish the asymptotic behavior of $\hat{S}_N$, we need to impose some further assumptions.

[4] Let $\mathbb{D}(b)$ be a generic stochastic process indexed by $b \in \mathcal{B}$. $\mathbb{D}(b)$ is said to be a process on $L^\infty(\mathcal{B})$ if $\mathbb{D}(\cdot)$ (treated as a random function with domain $\mathcal{B}$) has almost sure bounded paths (i.e., realizations) on $\mathcal{B}$.

Assumption 3.2 *The eigenvalues of* $\mathbb{E}\left[p^{k_N}(x_t)\, p^{k_N'}(x_t)\right]$ *for* $t = 1, .., T$ *are uniformly bounded and uniformly bounded away from zero.*

Assumption 3.3 Θ_N *is a closed subset of* Θ, *and there exists* $\Pi_N\theta \in \Theta_N$ *for each* $\theta \in \Theta$ *such that: (i)* $\delta_{s,N} \equiv \sup_{\theta\in\Theta\cap\Theta_R} \|\Pi_N\theta - \theta\|_{L^2} = o\left(N^{-1/4}\right)$*; (ii)* $\delta_{w,N} \equiv \sup_{\theta\in\Theta_I\cap\Theta_R} \mathrm{d}_w\left(\Pi_N\theta, \theta\right) = o\left(N^{-1/2}\right)$*; (iii) For some* $\gamma_N \downarrow 0$ *s.t.* $\sqrt{N}\gamma_N \uparrow \infty$, $\sup_{\theta\in\Theta_I\cap\Theta_R} \|\Pi_N\theta\|_s \leq B - \gamma_N$.

Assumption 3.4 $\varrho_N = O\left(N^{(1-\epsilon)/(4(2-\epsilon))}\right)$ *for arbitrarily small* $\epsilon > 0$.

Assumption 3.2 is identical to Assumption 3.3(i) in Santos (2012) and is commonly assumed in the literature on sieve estimation. Assumption 3.3(i) requires a uniform sieve approximation error rate under $\|\cdot\|_{L^2}$ over $\Theta\cap\Theta_R$. As discussed previously, the compactness of Θ under $\|\cdot\|_s$ makes the results in Schumaker (2007) applicable to developing a primitive condition for Assumption 3.3(i). Specifically, according to Theorem 6.25 in Schumaker (2007), $\sup_{\theta\in\Theta} \|\Pi_N\theta - \theta\|_{L^2} = O(k_N^{-(d-1)})$ for B-splines with simple knots. Therefore, Assumption 3.3(i) is satisfied by picking $k_N \to \infty$ fast enough such that $1/k_N = o\left(N^{-1/4(d-1)}\right)$. Since $\mathrm{d}_w\left(\Pi_N\theta, \theta\right)$ is controlled from above by $\|\Pi_N\theta - \theta\|_{L^2}$ as shown by Lemma 3.1, Assumption 3.3(ii) can be verified by using results for $\|\cdot\|_{L^2}$. Assumption 3.3(iii) is identical to Assumption 3.4(iii) in Santos (2012). Assumption 3.4 allows $\varrho_N \to \infty$, but restricts its divergence rate to be slow enough. Essentially, this requires us to pick k_N to grow sufficiently slow. Similar assumptions are commonly required in semi/non-parametric analyses for regularization of ill-posed problems. See, e.g., Blundell, Chen, and Kristensen (2007), Chen and Pouzo (2012), and Hong (2017). As acknowledged in Hong (2017), Assumption 3.4 is generally hard to verify because the nature of the dependence of ϱ_N on k_N has not been well studied. In the mildly ill-posed case as classified by Chen and Pouzo (2012), where $\varrho_N \asymp k_N^{\varpi}$ for some nonnegative constant ϖ, Assumption 3.4 is satisfied with $k_N = O\left(N^{(1-\epsilon)/(4(2-\epsilon)\varpi)}\right)$. Consequently, as long as we require $g(\cdot)$ to be sufficiently smooth such that $d \geq (2-\epsilon)\varpi/(1-\epsilon) + 1$, there exists k_N whose rate would satisfy both Assumptions 3.3 and 3.4. Also, in the special case of $R = 0$ (i.e., no IFEs) and point-identification, Assumption 3.2 is sufficient to guarantee $\mathrm{d}_w\left(\theta, \theta^0\right) \asymp \left\|\theta - \theta^0\right\|_{L^2}$ asymptotically for any $\theta^0 \in \Theta_I$ and $\left\|\theta - \theta^0\right\|_{L^2} = o(1)$, which implies that $\varrho_N = O(1)$. Then Assumption 3.4 holds trivially. In the online supplementary Appendix C, we clarify this claim for the case where $R = 0$ and also provide some further discussions on the sufficient conditions for Assumption 3.4 when $R \geq 1$.

With the above additional assumptions, we can state the next main result in this paper.

Theorem 3.2 (Consistency of $\hat{\theta}_N$) *Let Assumptions 2.1, 2.2, and 3.1 - 3.4 hold. For any* $\hat{\theta}_N \in \operatorname{argmin}_{\theta\in\Theta_N\cap\Theta_R} S_N(\theta)$, *it holds that*

$$\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) = O_p(\min(N^{-1/4}, \varrho_N N^{-1/2})) = o_p(N^{-\frac{1}{2}+\frac{\epsilon}{4}}\varrho_N^{1-\epsilon}). \tag{3.13}$$

A close examination of the proof of Theorem 3.2 suggests that we can first show that $\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) = O_p(N^{-1/4})$ under Assumptions 2.1, 2.2, and 3.1 - 3.3. Such a rate can be improved to $o_p(N^{-\frac{1}{2}+\frac{\epsilon}{4}}\varrho_N^{1-\epsilon})$ by using the link between d_w and $\mathrm{d}_{\|\cdot\|_{L^2}}$ through the sieve measure of ill-posedness and some iterative arguments. By Assumptions 3.3 - 3.4 and Lemma A.6 in Appendix A, we can show that $\mathrm{d}_{\|\cdot\|_{L^2}}\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) = o_p(\varrho_N^{2-\epsilon}N^{-\frac{1}{2}+\frac{\epsilon}{4}}) = o_p(N^{-1/4})$, which will be used in the proof of the next main result.

Theorem 3.3 (Asymptotic distribution under $\mathbb{H}_0$) *Let Assumption 2.1, 2.2, and 3.1 - 3.4 hold. Under the null hypothesis $\mathbb{H}_0 : \Theta_I \cap \Theta_R \neq \emptyset$, we have*

$$\hat{S}_N \xrightarrow{\mathcal{L}} \inf_{\theta \in \Theta_I \cap \Theta_R} \sum_{s=1}^{T-R} [\mathbb{B}_s(\theta) + \mathbb{C}_s(\theta)].$$

Theorem 3.3 studies the asymptotic distribution of $\hat{S}_N$ under the null hypothesis. Apparently, $\hat{S}_N$ is not asymptotically pivotal and we will provide a bootstrap method in the next subsection to obtain its bootstrap p-value for the purpose of inference.

Theorem 3.4 (Asymptotic behavior under $\mathbb{H}_1$) *Let Assumption 2.1, 2.2, and 3.1 - 3.4 hold. Under the alternative hypothesis $\mathbb{H}_1 : \Theta_I \cap \Theta_R = \emptyset$, we have*

$$N^{-1}\hat{S}_N \xrightarrow{p} \min_{\theta \in \Theta \cap \Theta_R} \sum_{s=1}^{T-R} \text{MDD}\left[m_s(Y, X, \theta) | \underline{z}_s\right]^2 > 0.$$

Theorem 3.4 studies the asymptotic behavior of $\hat{S}_N$ under the alternative. It indicates that $\hat{S}_N$ diverges to infinity in probability at rate$-N$, which gives the power of the MDD-based test.

3.4 A multiplier bootstrap

As shown by Theorem 3.3, the asymptotic distribution under the null hypothesis is nonstandard and unfamiliar. Here we propose a bootstrap procedure to obtain the bootstrap p-values. To ensure the consistency of the bootstrap procedure, we need to find appropriate ways to: (i) mimic the limiting law of $\hat{S}_N$ that is associated with a Gaussian chaos process $\mathbb{C}(\theta) = (\mathbb{C}_1(\theta), ..., \mathbb{C}_{T-R}(\theta))'$ on $L^\infty(\Theta_I)$ under the null, and (ii) ensure the bootstrap statistic is well behaved or divergent to infinity at a slower rate than $\hat{S}_N$ under the global alternative.

Let $\{v_i\}_{i=1}^N$ be an i.i.d. sequence that has mean zero and variance 1 and that is independent of the sample $\{(Y_i, X_i, Z_i)\}_{i=1}^N$. Two popular choices of distributions for $\{v_i\}_{i=1}^N$ are given by the standard normal distribution ($N(0,1)$) and the two-point distribution:

$$v_i = \begin{cases} -\left(\sqrt{5}-1\right)/2 & \text{with prob. } \left(\sqrt{5}+1\right)/\left(2\sqrt{5}\right) \\ \left(\sqrt{5}+1\right)/2 & \text{with prob. } \left(\sqrt{5}-1\right)/\left(2\sqrt{5}\right) \end{cases}. \tag{3.14}$$

Let $m_{is}^*(\theta) \equiv m_s(Y_i, \phi_s, \mathbf{g}(X_i)) v_i$. Motivated by the idea of multiplier bootstrap that is widely used for statistic tests involved with empirical processes or non-degenerate U-processes, we consider the following process

$$S_N^*(\theta) \equiv \sum_{s=1}^{T-R} S_{Ns}^*(\theta), \tag{3.15}$$

where

$$S_{Ns}^*(\theta) \equiv -\frac{1}{N} \sum_{1 \leq i \neq j \leq N} m_{is}^*(\theta) m_{js}^*(\theta) \kappa_{ij,s} + \frac{2}{N} \sum_{1 \leq i \neq j \leq N} m_{is}^*(\theta) \kappa_{ij,s} \frac{1}{N} \sum_{k=1}^N m_{ks}^*(\theta). \tag{3.16}$$

Let P^* and $\mathbb{E}^*$ denote respectively the probability law and expectation associated with (Y_i, X_i, Z_i, v_i) in the bootstrap world. We make two remarks on the construction of $S_N^*(\theta)$. First, note that we perturb

$m_s(Y_i, \phi_s, \mathbf{g}(X_i))$ through the multiplication by the random variable v_i that ensures $\mathbb{E}^*[m^*_{is}(\theta)] = 0$. This ensures that the dominant random component in the process $\{S^*_{Ns}(\theta)\}$ is given by a degenerate second order U-process that converges to a Gaussian chaos process. More importantly, we can show that the limiting law of $\{S^*_{Ns}(\theta)\}$ coincides with that of $\{S_{Ns}(\theta)\}$ on Θ_I. Second, $\{S^*_{Ns}(\theta)\}$ is also well behaved for $\theta \in \Theta \backslash \Theta_I$ (i.e., it is not divergent on $\Theta\backslash\Theta_I$), and if we were to define the bootstrap statistic as $\min_{\theta \in \Theta_N \cap \Theta_R} S^*_N(\theta)$, there is no way to ensure that the minimum is achieved at some value in $\Theta_I \cap \Theta_R$ asymptotically. In order to obtain the same limiting law for the bootstrap test statistic as $\hat{S}_N$ under the null, we must ensure that the minimum is achieved in the bootstrap world for some value of θ in $\Theta_I \cap \Theta_R$ when the null hypothesis holds true. Fortunately, this can be achieved be adding a suitable penalty term to the bootstrap minimization objective function, yielding the following bootstrap test statistic:

$$\hat{S}^*_N = \min_{\theta \in \Theta_N \cap \Theta_R} \left[S^*_N(\theta) + \mu_N \frac{S_N(\theta)}{N} \right],$$

where $P_N(\theta) \equiv \frac{1}{N} S_N(\theta)$ is a penalty term that ensures the minimum is achieved asymptotically for $\theta \in \Theta_I \cap \Theta_R$ under the null, and μ_N is a tuning parameter that diverges to infinity at a suitable rate (see Assumption 3.5 below). As a result, $\hat{S}^*_N$ shares the same limiting distribution as $\hat{S}_N$ under the null. This ensures the first goal mentioned above.

To ensure the good power properties of the bootstrap test, we require that $\hat{S}^*_N$ be well behaved under the alternative. When μ_N diverges to infinity at a rate slower than $N/\log(\log(N))$, we will show that $\mu_N^{-1} \hat{S}^*_N \xrightarrow{p} \min_{\theta \in \Theta \cap \Theta_R} \sum_{s=1}^{T-R} \mathbb{E}\left[\{\text{MDD}[m_s(Y, X, \theta) | \underline{z}_s]\}^2\right]$ under $\mathbb{H}_1 : \Theta_I \cap \Theta_R = \emptyset$. That is, $\hat{S}^*_N$ diverges to infinity at rate μ_N, which is slower than the rate N at which $\hat{S}_N$ diverges to infinity under the alternative. This implies that $\hat{S}_N \gg \hat{S}^*_N$ with probability approaching one (w.p.a.1) under the alternative, ensuring the second aforementioned goal.

To proceed, we add the following assumption on $\{v_i\}_{i=1}^N$ and the tuning parameter μ_N.

Assumption 3.5 *(i) $\{v_i\}_{i=1}^N$ is i.i.d. with mean zero and variance one, and is independent of $\{(Y_i, X_i, Z_i)\}_{i=1}^N$.*
(ii) As $N \to \infty$, $\mu_N \to \infty$ and $\mu_N = o(N/\log(\log(N)))$.

The following theorem states the asymptotic properties of $\hat{S}^*_N$ when the null hypothesis holds true or is violated.

Theorem 3.5 (Consistency of the multiplier bootstrap) *Let Assumption 2.1, 2.2, and 3.1 - 3.5 hold. If $\mathbb{H}_0 : \Theta_I \cap \Theta_R \neq \emptyset$ holds true, then*

$$\hat{S}^*_N \xrightarrow{\mathcal{L}} \inf_{\theta \in \Theta_I \cap \Theta_R} \sum_{s=1}^{T-R} [\mathbb{B}_s(\theta) + \mathbb{C}_s(\theta)];$$

And if $\mathbb{H}_1 : \Theta_I \cap \Theta_R = \emptyset$ holds true, then

$$\mu_N^{-1} \hat{S}^*_N \xrightarrow{p} \min_{\theta \in \Theta \cap \Theta_R} \sum_{s=1}^{T-R} \text{MDD}[m_s(Y, X, \theta) | \underline{z}_s]^2.$$

Remark. Theorem 3.5 shows that, with a properly chosen sequence of $\{\mu_N\}$, $\hat{S}^*_N$ converges weakly to the same asymptotic distribution as the original test statistic $\hat{S}_N$ does under the null hypothesis. This ensures the asymptotic level of the multiplier-bootstrap-based test. Theorem 3.5 also shows that, under any fixed alternative, $\hat{S}_N^*$ diverges to infinity at rate-μ_N, which is slower than rate-N at which $\hat{S}_N$ diverges to infinity.[5] Therefore, the proposed bootstrap procedure has asymptotic power one against any fixed alternative.

An essential step in the proof of Theorem 3.5 is the study of the unconditional central limit theorem (CLT) of $S_N^*(\theta)$, which is analogous to the unconditional multiplier CLT for empirical processes studied in van der Vaart and Wellner (1996, pp.177-181) and Kororok (2008, pp.181-183). It extends the unconditional multiplier CLT for degenerate second order U-statistics in Leucht and Neumann (2013) to degenerate second order U-processes. A close examination of the proof suggests that $\hat{S}_N^*$ is asymptotically independent of $\hat{S}_N$ under the null and different $\hat{S}_N^*$'s based on different independent sequences $\{v_i\}_{i=1}^N$ are also asymptotically independent of each other. This suggests that in practice, we can draw $\{v_i\}_{i=1}^N$ B times independently from suitable distributions to construct B bootstrap test statistics $\left\{\hat{S}_N^{*(b)}\right\}_{b=1}^B$. Then we can calculate the bootstrap p-value for our test statistic $\hat{S}_N$ as $p^* = \frac{1}{B}\sum_{b=1}^B \mathbf{1}\left\{\hat{S}_N \leq \hat{S}_N^{*(b)}\right\}$ with $\mathbf{1}\{\cdot\}$ being the usual indicator function, and reject the null hypothesis when p^* is smaller than the prescribed level of significance.

4 Monte Carlo Simulations

In this section, we conduct Monte Carlo simulations to evaluate the finite sample performance of our proposed inference method.

4.1 Design 1

First, we consider a data generating process (DGP) similar to the one used by Santos (2012) in his Monte Carlo study, with an added IFE term. We set $T = 2$ and $R = 1$. For each period $t = 1, 2$, we generate an i.i.d. sample by

$$\begin{pmatrix} x_{it}^\blacklozenge \\ z_{it}^\blacklozenge \\ u_{it}^\blacklozenge \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} 1.0 & 0.5 & 0.3 \\ 0.5 & 1.0 & 0.0 \\ 0.3 & 0.0 & 0.5 \end{bmatrix}\right).$$

Then we use the latent variables $\left(x_t^\blacklozenge, z_t^\blacklozenge, u_t^\blacklozenge\right)$ to generate (x_t, z_t, u_t) as follows:

$$x_{it} = 2\left[\Psi\left(x_{it}^\blacklozenge/3\right) - 0.5\right],\ z_{it} = 2\left[\Psi\left(z_{it}^\blacklozenge/3\right) - 0.5\right],\ \text{and } u_{it} = u_{it}^\blacklozenge,$$

where $\Psi(\cdot)$ is the CDF of the standard normal distribution. Finally, the dependent variable y_t is generated as

$$y_{it} = 2\cos(x_{it}\pi) - 2 + \lambda_i F_t + u_{it},$$

with $F' = \begin{pmatrix} F_1 & F_2 \end{pmatrix} = \begin{pmatrix} 0.1 & -1.0 \end{pmatrix}$ and $\lambda_i \sim$ i.i.d. $\mathrm{N}(0, 0.5)$. Here we generate $\{\lambda_i\}$ independently of all other variables.

[5] Recall that if $\mathbb{H}_1 : \Theta_I \cap \Theta_R = \emptyset$ holds true, then $\min_{\theta \in \Theta \cap \Theta_R} \sum_{s=1}^{T-R} \mathrm{MDD}\left[m_s(Y, X, \theta) | \underline{z}_s\right]^2 > 0$, which has already been stated in Theorem 3.4.

Table 1: Size performance for Desgin 1

		$N=250$				$N=500$			
		$B=5$		$B=15$		$B=5$		$B=15$	
α	μ_N	2-pt	Norm	2-pt	Norm	2-pt	Norm	2-pt	Norm
0.10	0	0.172	0.168	0.196	0.184	0.224	0.216	0.236	0.252
0.10	$N^{1/4}$	0.118	0.122	0.146	0.148	0.132	0.114	0.164	0.162
0.10	$N^{1/3}$	0.092	0.100	0.126	0.130	0.096	0.102	0.132	0.140
0.10	$N^{1/2}$	0.064	0.060	0.082	0.092	0.046	0.044	0.068	0.064
0.05	0	0.090	0.086	0.134	0.128	0.106	0.106	0.132	0.144
0.05	$N^{1/4}$	0.072	0.070	0.098	0.092	0.058	0.062	0.096	0.098
0.05	$N^{1/3}$	0.058	0.062	0.088	0.086	0.044	0.046	0.078	0.060
0.05	$N^{1/2}$	0.038	0.040	0.060	0.062	0.018	0.018	0.038	0.032
0.01	0	0.028	0.026	0.060	0.042	0.014	0.016	0.034	0.036
0.01	$N^{1/4}$	0.024	0.020	0.046	0.050	0.010	0.012	0.030	0.028
0.01	$N^{1/3}$	0.022	0.020	0.038	0.048	0.002	0.010	0.020	0.024
0.01	$N^{1/2}$	0.012	0.008	0.030	0.026	0.000	0.000	0.018	0.014

Note: α is the nominal size; B is the norm upper bound; '2-pt' and 'Norm' refer to two-point and standard normal distributions, respectively.

We primarily focus on studying the effect of the penalty weight μ_N and the norm constraint B on the empirical size/level of the test. For this purpose, we consider the family of null hypotheses

$$\mathbb{H}_0 : \Theta_I \cap \Theta_{R(\gamma)} \neq \emptyset, \quad \Theta_{R(\gamma)} \equiv \{\theta : g'(0) = \gamma\}. \tag{4.1}$$

Under the DGP specified above, the model is nonparametrically identified with a location normalizing condition $g(0) = 0$. And the null hypothesis in (4.1) is true only at $\gamma = 0$, and is false otherwise. We set $d = 3$, which satisfies the requirement $d \geq d_x + 2$ of Assumption 2.1(i). The sieve was chosen to be B-spline of order 3 with knots $\{-1, -1, -1, 0, 1, 1, 1\}$, which implies $k_N = 4$. We conduct 200 bootstrap evaluations to calculate the bootstrap p-values. The Monte Carlo study consists of 500 replications.

In Table 1, we report the simulated size under Design 1 for testing the null hypothesis in (4.1) with $\gamma = 0$, as a function of the targeted size α, norm constraint B, penalty weight μ_N, and sample size N, for two different disturbance distributions used in the multiplier bootstrap procedure (the two-point distribution specified by (3.14) and the standard normal distribution). As mentioned above, the null hypothesis in (4.1) is true at $\gamma = 0$ under Design 1. From Table 1 we have the following observations: (i) The choice $\mu_N = 0$ (i.e., no penalty in the bootstraps) leads to severe size distortions, as expected from our theory, and is considered to illustrate the extreme case of selecting too small a penalty weight; (ii) The size is not sensitive to either the choice of disturbance distribution or the choice of B; (iii) The size is somewhat sensitive to the choice of μ_N. Overall, choosing $\mu_N = N^{1/3}$ gives good size control under the current design.

In Table 2, we report the simulated probability of rejecting the null hypothesis in (4.1) at $\gamma = \pm 1.3$, ± 1.9, and ±3.1, under a nominal size of 0.05 and Design 1. Since the null hypothesis in (4.1) is false at any of these γ values, Table 2 means to assess the finite sample power performance against fixed alternatives. Note that the further γ is away from zero, the further the hypothesis in (4.1) deviates from a true one. From Table 2 we have the following observations: (i) Overall, the rejection probabilities are noticeably larger than the nominal size 0.05, and they keep getting larger as γ moves away from zero. This indicates good power performance; (ii) For almost all given values of γ and given choices of μ_N, B, and disturbance distributions used in bootstraps, the rejection probabilities generally increase as we increase the sample size from $N = 250$ to $N = 500$;[6] (iii) The powers are not sensitive to the choice of disturbance distribution; (iv) The powers are somewhat sensitive to the choice of μ_N and B. With a smaller B and a smaller μ_N, we get higher rejection probabilities. Nevertheless, the overall results indicates good power performance, as noted in (i).

4.2 Design 2

As a robustness check, we also conduct Monte Carlo simulations under a DGP that is different from Design 1. Now we allow the IFEs to be correlated with the covariate x_{it}. Specifically, now we generate x_{it} as $x_{it} = 2\left[\Psi\left((x_{it}^{\blacklozenge} + \lambda_i' F_t)/3\right) - 0.5\right]$ while keeping everything else unchanged from the previous DGP. We still consider the null hypotheses specified in (4.1). Under the current DGP, the model is nonparametrically identified. And the null in (4.1) is true only at $\gamma = 0$, and is false otherwise. Like in Design 1, we conduct 200 bootstrap evaluations to calculate the bootstrap p-values. The Monte Carlo study consists of 500 replications. Under Design 2, we report simulated probabilities of rejecting the null at $\gamma = 0$ in Table 3. And we report simulated rejection probabilities at $\gamma = \pm 1.3$, ± 1.9, and ± 3.1, under a nominal size of 0.05, in Table 4. From these results, we obtain observations very similar to those obtained from Tables 1 and 2.

4.3 Design 3

Now we focus on testing for linearity. We test the null hypothesis

$$\mathbb{H}_0 : \Theta_I \cap \Theta_{R_L} \neq \emptyset, \quad \Theta_{R_L} \equiv \left\{\theta \in \Theta : g(x) = a + bx \text{ for some } (a, b) \in \mathbb{R}^2\right\}. \tag{4.2}$$

And we adopt a series of DGPs indexed by γ with $T = 3$ and $R = 1$, as follows:

$$\begin{aligned} y_{it} &= x_{it} + \gamma x_{it}^2 + \lambda_i F_t + u_{it} \\ x_{it} &= 0.25\lambda_i' F_t + 0.5 z_{it} + 0.5 u_{it} + \varepsilon_{it} \end{aligned}$$

with

$$\begin{pmatrix} z_{it} \\ \varepsilon_{it} \\ u_{it} \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.5 \end{bmatrix}\right),$$

[6] There are only few exceptions found in Table 2, where the rejection probabilities do not increase as we increase the sample size from $N = 250$ to $N = 500$. They all happen when setting $B = 15$, with $\{\gamma = -1.3,\ \mu_N = N^{1/4},\ \text{Two-pt}\}$, $\{\gamma = -1.3,\ \mu_N = N^{1/3},\ \text{Normal}\}$, $\{\gamma = -1.3,\ \mu_N = N^{1/4},\ \text{Normal}\}$, $\{\gamma = -1.3,\ \mu_N = N^{1/2},\ \text{Normal}\}$, and $\{\gamma = -1.9,\ \mu_N = N^{1/2},\ \text{Two-pt}\}$, respectively. For the former two, the rejection probabilities decrease slightly. For the latter three, the rejection probabilities stay the same.

Table 2: Power performance for Desgin 1 (nominal size: 0.05)

		$N = 250$				$N = 500$			
		$B = 5$		$B = 15$		$B = 5$		$B = 15$	
γ	μ_N	2-pt	Norm	2-pt	Norm	2-pt	Norm	2-pt	Norm
1.3	$N^{1/4}$	0.250	0.244	0.154	0.140	0.444	0.442	0.182	0.178
1.3	$N^{1/3}$	0.230	0.234	0.138	0.140	0.410	0.392	0.142	0.138
1.3	$N^{1/2}$	0.150	0.150	0.080	0.088	0.262	0.264	0.074	0.074
-1.3	$N^{1/4}$	0.194	0.198	0.098	0.090	0.314	0.330	0.082	0.090
-1.3	$N^{1/3}$	0.174	0.174	0.068	0.072	0.278	0.260	0.074	0.068
-1.3	$N^{1/2}$	0.106	0.114	0.056	0.058	0.182	0.182	0.060	0.058
1.9	$N^{1/4}$	0.534	0.536	0.240	0.242	0.826	0.820	0.338	0.340
1.9	$N^{1/3}$	0.498	0.488	0.224	0.210	0.786	0.782	0.284	0.310
1.9	$N^{1/2}$	0.390	0.386	0.146	0.146	0.670	0.646	0.150	0.154
-1.9	$N^{1/4}$	0.360	0.370	0.140	0.142	0.650	0.644	0.182	0.166
-1.9	$N^{1/3}$	0.346	0.350	0.118	0.128	0.602	0.610	0.152	0.150
-1.9	$N^{1/2}$	0.258	0.260	0.088	0.082	0.490	0.492	0.088	0.086
3.1	$N^{1/4}$	0.960	0.952	0.714	0.714	0.998	0.998	0.934	0.922
3.1	$N^{1/3}$	0.942	0.936	0.680	0.676	0.998	0.998	0.912	0.922
3.1	$N^{1/2}$	0.898	0.908	0.552	0.522	0.994	0.998	0.818	0.820
-3.1	$N^{1/4}$	0.796	0.792	0.438	0.432	0.972	0.974	0.746	0.740
-3.1	$N^{1/3}$	0.776	0.768	0.416	0.418	0.968	0.964	0.708	0.698
-3.1	$N^{1/2}$	0.680	0.662	0.352	0.348	0.948	0.942	0.606	0.602

Note: B is the norm upper bound; '2-pt' and 'Norm' refer to two-point and standard normal distributions, respectively.

Table 3: Size performance for Desgin 2

		$N = 250$				$N = 500$			
		$B = 5$		$B = 15$		$B = 5$		$B = 15$	
α	μ_N	2-pt	Norm	2-pt	Norm	2-pt	Norm	2-pt	Norm
0.10	0	0.162	0.156	0.160	0.184	0.224	0.212	0.210	0.216
0.10	$N^{1/4}$	0.090	0.090	0.126	0.118	0.120	0.126	0.148	0.152
0.10	$N^{1/3}$	0.060	0.060	0.100	0.102	0.098	0.094	0.134	0.128
0.10	$N^{1/2}$	0.032	0.038	0.052	0.048	0.044	0.054	0.082	0.008
0.05	0	0.068	0.066	0.094	0.098	0.104	0.102	0.132	0.140
0.05	$N^{1/4}$	0.038	0.038	0.058	0.056	0.052	0.058	0.108	0.102
0.05	$N^{1/3}$	0.036	0.034	0.050	0.048	0.046	0.044	0.086	0.086
0.05	$N^{1/2}$	0.018	0.016	0.024	0.024	0.012	0.018	0.046	0.048
0.01	0	0.010	0.016	0.018	0.024	0.014	0.018	0.046	0.046
0.01	$N^{1/4}$	0.012	0.014	0.014	0.018	0.004	0.006	0.034	0.040
0.01	$N^{1/3}$	0.010	0.012	0.014	0.012	0.004	0.004	0.026	0.026
0.01	$N^{1/2}$	0.004	0.008	0.010	0.008	0.002	0.002	0.016	0.018

Note: α is the nominal size; B is the norm upper bound; '2-pt' and 'Norm' refer to two-point and standard normal distributions, respectively.

Table 4: Power performance for Desgin 2 (nominal size: 0.05)

		$N = 250$				$N = 500$			
		$B = 5$		$B = 15$		$B = 5$		$B = 15$	
γ	μ_N	2-pt	Norm	2-pt	Norm	2-pt	Norm	2-pt	Norm
1.3	$N^{1/4}$	0.260	0.262	0.158	0.140	0.478	0.482	0.170	0.188
1.3	$N^{1/3}$	0.230	0.230	0.114	0.130	0.422	0.434	0.150	0.146
1.3	$N^{1/2}$	0.152	0.152	0.050	0.062	0.286	0.280	0.080	0.074
-1.3	$N^{1/4}$	0.162	0.160	0.058	0.062	0.274	0.258	0.084	0.076
-1.3	$N^{1/3}$	0.132	0.128	0.056	0.058	0.238	0.228	0.066	0.068
-1.3	$N^{1/2}$	0.076	0.074	0.052	0.054	0.154	0.156	0.052	0.052
1.9	$N^{1/4}$	0.550	0.552	0.252	0.258	0.874	0.880	0.368	0.364
1.9	$N^{1/3}$	0.510	0.514	0.226	0.232	0.850	0.848	0.294	0.316
1.9	$N^{1/2}$	0.406	0.418	0.138	0.142	0.742	0.746	0.174	0.186
-1.9	$N^{1/4}$	0.330	0.346	0.102	0.106	0.624	0.610	0.146	0.146
-1.9	$N^{1/3}$	0.296	0.302	0.086	0.094	0.600	0.578	0.114	0.116
-1.9	$N^{1/2}$	0.226	0.202	0.060	0.066	0.458	0.448	0.072	0.066
3.1	$N^{1/4}$	0.926	0.932	0.696	0.700	1.000	1.000	0.944	0.950
3.1	$N^{1/3}$	0.914	0.916	0.680	0.676	1.000	1.000	0.936	0.930
3.1	$N^{1/2}$	0.886	0.884	0.574	0.560	1.000	1.000	0.874	0.880
-3.1	$N^{1/4}$	0.764	0.746	0.420	0.426	0.970	0.974	0.712	0.696
-3.1	$N^{1/3}$	0.736	0.720	0.404	0.378	0.966	0.972	0.684	0.682
-3.1	$N^{1/2}$	0.628	0.626	0.302	0.294	0.936	0.938	0.562	0.562

Note: B is the norm upper bound; '2-pt' and 'Norm' refer to two-point and standard normal distributions, respectively.

Table 5: Size and power performance for Design 3 (nominal size: 0.05)

γ	μ_N	$N = 250$	$N = 500$	$N = 1000$
0	$N^{1/4}$	0.012	0.026	0.044
0	$N^{1/3}$	0.010	0.020	0.040
0	$N^{1/2}$	0.006	0.016	0.028
0.2	$N^{1/4}$	0.046	0.156	0.296
0.2	$N^{1/3}$	0.038	0.150	0.272
0.2	$N^{1/2}$	0.016	0.102	0.188
0.5	$N^{1/4}$	0.112	0.410	0.848
0.5	$N^{1/3}$	0.106	0.392	0.832
0.5	$N^{1/2}$	0.068	0.290	0.788
1.0	$N^{1/4}$	0.232	0.570	0.956
1.0	$N^{1/3}$	0.204	0.526	0.940
1.0	$N^{1/2}$	0.126	0.454	0.908

Note: Rows with $\gamma = 0$ show size and all other rows show power.

$F' = \begin{pmatrix} F_1 & F_2 & F_3 \end{pmatrix} = \begin{pmatrix} 0.7 & 0.2 & -1.0 \end{pmatrix}$, and $\lambda_i \sim$ i.i.d. $\mathrm{N}(0, 0.1)$. Again, we generate $\{\lambda_i\}$ independently of all other variables.

For a given γ, denote the above DGP as DGP(γ). Under DGP(γ), the null in (4.2) is true or false, depending on the value of γ. The null is true under DGP(0), and is false under DGP(γ) at any given $\gamma \neq 0$. And γ can be viewed as a measure of how far the DGP is away from a linear specification. We do 200 bootstrap evaluations using the two-point distribution and adopt 500 replications. Table 5 reports the simulated probabilities of rejecting the null in (4.2) under a nominal size of 0.05 for DGP(γ) with $\gamma = 0$, 0.2, 0.5 and 1, and for $N = 250$, 500 and 1000. Note that rows with $\gamma = 0$ show size, while all other rows show power. From Table 5 we have the following observations: (i) $N = 250$ seems to be too small for this design; (ii) As we increase the sample size, the simulated size keeps getting closer to the targeted size of 0.05. When $N = 1000$, the simulated size is reasonably close to 0.05 with $\mu_N = N^{1/4}$ and $N^{1/3}$; (iii) The rejection probabilities increase noticeably as γ deviates away from zero, for all choices of μ_N and all sample sizes. This indicates good power performance. (iv) Similar to what we observe in the previous two designs, $\mu_N = N^{1/2}$ seems to be a bit too large for the penalty weight, which leads to undersizing even at $N = 1000$ for the current design. Nevertheless, $\mu_N = N^{1/2}$ still provides good power performance.

5 Empirical Application

In this section, we apply our method to study Engel curves for major nondurable expenditures in China, using data from the China Family Panel Studies (CFPS) for the period 2010 - 2014. The CFPS is similar to the U.K. Family Expenditure Survey (U.K.FES), but is conducted every other year. Specifically, the CFPS data are collected in 2010, 2012, and 2014, which produces a three-period ($T = 3$) balanced panel data set of 6627 households.

According to the panel data from CFPS and its consumption categorization, on average, food (including dining) expenditures take the largest share of total nondurable expenditures (averaging at 41.54%), which is followed by medical and health care expenditures (averaging at 12.01%), expenditures on commute and communication (averaging at 9.99%), and grocery expenditures (averaging at 9.63%). We study the Engel curves for these four major categories of consumption.

For household i at period t, let $y_{f,it}$, $y_{m,it}$, $y_{c,it}$, $y_{g,it}$ be the share of its total nondurable expenditures spent on food including dining (FD), medical and health care (MH), commute and communication (CC), grocery(GY), respectively. Let x_{it} be the log of total annual nondurable expenditures, and let z_{it} be total household annual income. The Engel curves are assumed to take the following additive form:

$$y_{j,it} = g_j\left(x_{it}\right) + \lambda'_{j,i} F_{j,t} + u_{j,it}, \tag{5.1}$$

for $t \in \{1, 2, 3\}$ and $j \in \{f, m, c, g\}$. $F_{j,t}$, $\lambda_{j,i}$, and $u_{j,it}$ are unobservable terms, representing the vector of factors, the vector of factor loadings, and other heterogeneity, respectively. Some observations are dropped because they have key variables out of reasonable range.[7] For studying Engel curves for FD and MH, we also condition on households who exhibited positive consumption of both categories. These restrictions yield a three-period panel of 3811 cross-sectional observations for FD and MH (consisting of 1787 urban households and 2024 rural households). Similarly, for studying Engel curves for CC and GY, we condition on households who exhibited positive consumption of these two categories both, which yields a three-period panel of 4584 cross-sectional observations for CC and GY (consisting of 2287 urban households and 2297 rural households).

We set the support for x_{it} as $\mathcal{X} = [7, 14]$, which includes all observations for all four categories of consumption. Since $\mathcal{X}$ is compact, we set $\zeta = 0$ (i.e., no tail control needed). For nonparametric inferences, we employ B-spline of order 3 on $[7, 14]$, with 2 interior knots which are the 33.3th percentile and the 66.7th percentile of the corresponding x_{it} sample. And we specify $B = 15$. Finally, we conduct hypothesis tests with 200 bootstrap repetitions, and with $\mu_N = N^{1/4}$ and $\mu_N = N^{1/3}$ as suggested by our Monte Carlo simulations.

5.1 Testing for log-linearity

Since the seminar work of Deaton and Muellbauer (1980), a log-linear specification (i.e., linear in the log of total nondurable expenditures) has been commonly adopted to parameterize Engel curves in the literature. The use of log-linear Engel curves to estimate and correct bias in directly measured macroeconomic indicators is very prevalent in empirical studies. For instance, a popular way to construct an alternative measure of household income or expenditure relies on the log-linear Engel curves as a key assumption to infer incomes or expenditures. See Aguiar and Bils (2015), Browning and Crossley (2009), Hurst, Li, and Pugsley (2014), and Pissarides and Weber (1989), among others. Besides, there are papers focusing on estimating CPI bias

[7] We drop observations with extremely low total annual expenditures ($\leq$ 2188 CNY, which corresponds to the 0.01 quantile of total expenditures distribution) or extremely low total household annual income ($\leq$ 3000 CNY, which equals the poverty line per capita set by the State Council Leading Group Office of Poverty Alleviation and Development of China in 2015)

based on the log-linear form of Engel curves; see Hamilton (2001) and Nakamura, Steinsson, and Liu (2016), among others. On the other hand, there are also papers advocating advantages of building nonparametric Engel curves over the parametric ones for studies of demand. See, e.g., Blundell, Browning and Crawford (2003) and Blundell, Chen, and Kristensen (2007).

In our empirical study, we first examine whether the log-linear relationship could adequately describe the Engel curves for major nondurable expenditures in China. Under a potential lack of point-identification, the linear specification can be tested through the hypothesis specified in (4.2) that we previously studied in our Monte Carlo simulations, i.e.,

$$\mathbb{H}_0 : \Theta_I \cap \Theta_{R_L} \neq \emptyset, \quad \Theta_{R_L} \equiv \left\{\theta \in \Theta : g\left(x\right) = a + bx \text{ for some } (a,b) \in \mathbb{R}^2\right\}. \tag{5.2}$$

To account for potential differences in consumption pattern/habit between urban and rural households, we conduct tests of the above hypotheses using the whole sample, the urban subsample, and the rural subsample, respectively. In Table 6, we report the bootstrap p-values of corresponding test statistics for each consumption category, and for specifying the number of factors as $R = 1$ or 2. According to Table 6, when setting $R = 1$ and applying the conventional 5% significance level, we obtain the following testing results: (i) For FD, our tests reject the null except for the rural subsample; (ii) For MH, interestingly, our tests fail to reject the null when conditioning on either urban or rural households, yet are able to reject the null for the whole sample (i.e., without conditioning on urban or rural households); (iii) For both CC and GY, our tests reject the null regardless of whether we condition on urban households, rural households, or not. We get the same testing results regardless of whether $\mu_N = N^{1/4}$ or $N^{1/3}$, and the p-values seem to be insensitive to these two choices of μ_N. In short, these results suggest nonlinearity for some Engel curves and noticeable difference in consumption pattern/habit on FD and MH between urban and rural households.

Also according to Table 6, when setting $R = 2$, however, we fail to reject the null of log-linear specification for the Engel curves for all cases under our investigation. A possible explanation is that the IFE terms encompass unobserved heterogeneity of more flexible forms under $R = 2$ than that under $R = 1$; and consequently, under $R = 2$, the log-linear relationship might suffice to adequately describe the Engel curves, with heterogeneity being more flexibly taken care of by the IFEs. As noted by Santos (2012), failing to reject the null that there are log-linear Engel curves in the identified set does not necessarily justify adopting such a parametric specification. If the model is partially identified, then even when log-linear specifications are indeed in Θ_I, there is no guarantee that the true model is one of them. Therefore confidence intervals constructed under the log-linear assumption may asymptotically exclude the true parameter of interest.

5.2 Confidence interval for $g\left(\bar{x}\right)$

Next, we examine the robustness of our testing results regarding the log-linear specification by comparing 95% confidence intervals for food Engel curves at the sample average (across both N and T) $\bar{x}$ with and without assuming log-linearity and by setting $R = 1$ or 2 and $\mu_N = N^{1/3}$. Define

$$\Theta_{R_\gamma} \equiv \left\{\theta \in \Theta : g\left(\bar{x}\right) = \gamma\right\}. \tag{5.3}$$

Table 6: Bootstrap p-values for testing the log-linear null for the Engel curves

		NO. of factors $R = 1$			NO. of factors $R = 2$		
Cat.	μ_N	all	urban	rural	all	urban	rural
FD	$N^{1/4}$	0.000	0.000	0.915	0.520	0.460	0.930
MH	$N^{1/4}$	0.010	0.655	0.165	0.260	0.740	0.255
CC	$N^{1/4}$	0.000	0.000	0.000	0.060	0.415	0.185
GY	$N^{1/4}$	0.000	0.010	0.010	0.790	0.860	0.810
FD	$N^{1/3}$	0.005	0.000	0.865	0.635	0.495	0.960
MH	$N^{1/3}$	0.005	0.745	0.145	0.240	0.795	0.400
CC	$N^{1/3}$	0.000	0.000	0.000	0.075	0.460	0.230
GY	$N^{1/3}$	0.000	0.030	0.010	0.825	0.875	0.790

Note: FD, MH, CC and GY abbreviate food, medical and heath care, commute and communication, and grocery, respectively.

We obtain the confidence intervals for $g(\bar{x})$ under the log-linear specification by inverting tests of a series of null hypotheses, indexed by γ, defined as follows:

$$\mathbb{H}_{0,\gamma} : \Theta_I \cap \left(\Theta_{R_L} \cap \Theta_{R_\gamma}\right) \neq \emptyset, \tag{5.4}$$

where Θ_{R_L} is defined in (5.2). We also obtain the nonparametric confidence intervals $g(\bar{x})$ by inverting tests of the following null hypotheses:

$$\mathbb{H}_{0,\gamma} : \Theta_I \cap \Theta_{R_\gamma} \neq \emptyset, \tag{5.5}$$

where Θ_{R_γ} is defined by (5.3). For comparison purpose, we also construct confidence intervals using the method developed by Santos (2012), as well as standard IV confidence intervals. To make Santos' (2012) method applicable here, we pool the panel into a large cross-section data set, ignoring any potential fixed effects (which effectively turns into a situation with $T = 1$ and $R = 0$). Standard IV confidence intervals are constructed based on the pooled data set, too. To make a direct comparison between our method and Santos' (2012), we also construct confidence intervals based on the pooled data set using our method. All these confidence intervals are reported in Table 7. Here we mainly focus on results from the urban subsample for a detailed discussion. According to Table 7, for food consumption by urban households, while the log-linear confidence interval based on Santos (2012) is somewhat larger than the standard IV one, we end up with an empty set constructing log-linear confidence interval based on the pooled data set using our method. Moreover, based on the original panel data set, when setting $R < 2$, we obtain empty confidence intervals even under the nonparametric specification. Conducting further tests (to be discussed in the next subsection) on specifications using our method confirms our finding of these empty sets: for food consumption by urban households, based on the pooled data set, our test rejects the null of a log-linear specification; based on the original panel data set, our test reject both the null of a log-linear specification and that of a nonparametric specification when setting $R < 2$.[8] In the next subsection we focus on the joint specification of functional forms and the number of factors.

Interestingly, several confidence intervals are given by $[0, 1]$ in Table 7 under $R = 2$. This suggests that the inference does not provide any informative/powerful results regarding $g(\bar{x})$ when we allow for both a very flexible heterogeneity specification ($R = 2$) and a lack of point-identification. Overall, Table 7, in particular its first 6 rows, can be interpreted as results from a sensitivity analysis which demonstrates how the degree of informativeness/powerfulness of our inference procedure varies in response to changes in the strength of the model assumptions. These results pretty much reflect the law of decreasing credibility as coined by Manski (2003): Stronger assumptions yield inferences that may be more powerful but less credible, which is a dilemma faced by empirical researchers as they decide what assumption to maintain. Here, we also take Manski's (2003) view that statistical theory cannot resolve the dilemma but can clarify its nature. That said, Table 7 suggests our inference procedure still provides informative results under the already quite general setting of $R = 1$, nonparametric $g(\cdot)$ and partial identification.

5.3 Further investigation on heterogeneity and specification

We further investigate the specification of the functional form of $g(\cdot)$ and the number of factors (R), again focusing on food consumption. Ignoring the index j for different expenditure categories, we can rewrite the Engel curves in (5.1) as

$$y_{it} = g_{it}(x_{it}) + u_{it}, \tag{5.6}$$

$$g_{it}(x_{it}) = g(x_{it}) + \lambda_i' F_t, \tag{5.7}$$

where $t = 1, 2, 3$, the subscripts on $g_{it}(\cdot)$ capture potential heterogeneity cross individual and time, while (5.7) assumes the heterogeneity to take a special form of a common part $g(\cdot)$ augmented with an additive IFE term. A specification under the (5.6) - (5.7) framework is characterized by the combination of two specifications: (i) the functional form specification on $g(\cdot)$, and (ii) the specification on the number of factors R. The less restrictive a functional form specification on $g(\cdot)$, the more flexible/general the model is w.r.t. the common part relationship. The larger the number of factors, the more flexible/general the model is w.r.t. unobserved heterogeneity. It is easy to see that, within the (5.6) - (5.7) framework, the model achieves its maximum flexibility/generality when $g(\cdot)$ is treated nonparametrically and R is set to 2.[9] We use our method to test the specification of a variety these combinations.[10]

The p-values obtained from these tests are reported in Table 8. As suggested by Table 8, even with a nonparametric specification on $g(\cdot)$, the model does not suffice to adequately describe the Engel curve for food consumption among urban households in China when setting $R < 2$.[11] Interestingly, in comparison, our test fails to reject a log-linear specification on Engel curve among rural households when setting $R = 1$.

[8] When $R = 0$, one can continue to implement our testing procedure in the absence of the nuisance parameter ϕ.

[9] Recall that the maximum number of factors allowed is $T - 1$ for our method to work.

[10] In other words, here we view these tests as jointly testing for the specification on R and the specification on the functional form of $g(\cdot)$. While for obtaining CI for $g(\bar{x})$, we take a narrow view and interpret corresponding tests as only testing for the specification on the functional form of $g(\cdot)$, assuming any given specification on R to be true.

[11] As mentioned earlier in this section, we employ a specific B-spline (i.e., of order 3 on $[7, 14]$ with 2 interior knots) to

Table 7: 95% confidence intervals of $g(\bar{x})$ for the food Engel curves

	All	Urban	Rural
$R=2$, nonparametric	$[0.000, 1.000]$	$[0.000, 1.000]$	$[0.000, 1.000]$
$R=2$, log-linear	$[0.000, 1.000]$	$[0.110, 0.890]$	$[0.000, 1.000]$
$R=1$, nonparametric	Empty	Empty	$[0.308, 0.633]$
$R=1$, log-linear	Empty	Empty	$[0.390, 0.585]$
$R=0$, nonparametric	Empty	Empty	Empty
$R=0$, log-linear	Empty	Empty	Empty
Pooled, nonparametric	$[0.325, 0.530]$	$[0.387, 0.550]$	$[0.258, 0.511]$
Pooled, log-linear	$[0.406, 0.412]$	Empty	$[0.400, 0.410]$
Pooled, nonpara., Santos	$[0.119, 0.810]$	$[0.252, 0.712]$	$[0.000, 0.791]$
Pooled, log-linear, Santos	$[0.378, 0.440]$	$[0.366, 0.533]$	$[0.377, 0423]$
Pooled, standard IV	$[0.411, 0.418]$	$[0.419, 0.429]$	$[0.401, 0.411]$

Note: Given R, a log-linear CI for $g(\bar{x})$ is empty if the null $\mathbb{H}_{0,\gamma} : \Theta_I \cap \Theta_{R_L} \neq \emptyset$ is rejected at 5% level; Similarly, a nonparametric CI for $g(\bar{x})$ is emply if the null $\mathbb{H}_{0,\gamma} : \Theta_I \cap \Theta_R \neq \emptyset$ is rejected at 5% level.

[Note that our test still rejects a nonparametric specification for the rural population when setting $R = 0$.] In contrast, Banks, Blundell, and Lewbel (1997) pool a panel data set into a cross-sectional one and their study suggests that a log-quadratic specification suffices to adequately describe most Engel curves. Admittedly, such a comparison is only indirect because Banks, Blundell, and Lewbel (1997) use a different data set, one obtained from the U.K.FES, for their study. Our findings suggest: (i) There is a great degree of heterogeneity on food consumption pattern among urban households in China; (ii) There is a lesser degree of heterogeneity on food consumption pattern among rural households, compared with that among urban households in China; (iii) Even a nonparametric specification on $g(\cdot)$, as general as it is, might still be insufficient to compensate for an inadequate handling of heterogeneity to make the whole model a correctly specified one; (iv) When a panel data set is available, using methods that fully extract information from the panel structure, such as ours, could potentially provide more informative results than those obtained based on cross-sectional data sets, or based on panel data sets but treated as pooled cross-sectional ones.

approximate $g(\cdot)$ for corresponding tests where $g(\cdot)$ is supposed to be treated nonparametrically. This practice shares the same spirit with many existing nonparametric testing procedures. In an utterly strict sense, what is really tested here is where the specific cubic B-spline with 2 interior knots is adequate to describe the Engel curve for the given finite sample.

Table 8: P-values for testing joint specification on $g(\cdot)$ and R

	All	Urban	Rural
$R = 2$, nonparametric	0.750	0.960	0.975
$R = 2$, log-linear	0.635	0.495	0.960
$R = 1$, nonparametric	0.015	0.005	0.595
$R = 1$, log-linear	0.005	0.000	0.865
$R = 0$, nonparametric	0.000	0.000	0.000
$R = 0$, log-linear	0.000	0.000	0.000
Pooled, nonparametric	0.870	0.990	0.905
Pooled, log-linear	0.285	0.000	0.390
Pooled, nonpara., Santos	0.760	0.925	0.220
Pooled, log-linear, Santos	0.065	0.835	0.195

6 Conclusion

In this paper we propose a statistical inference procedure for partially identified nonparametric panel data models with endogeneity and IFEs. Even though the original identified set is specified through a set of conditional moment restrictions under the weak exogeneity assumption, we are able to translate it into an equivalent set of unconditional moment restrictions by using the novel MDD measure for the distance between a conditional mean object and zero. We construct the test statistic based on such a measure which is associated with a second order U-process in the limit that is degenerate under the null and non-degenerate under the alternative. We derive the limiting distribution of the resultant test statistic under the null and show that it is divergent at rate-N under the global alternative. To obtain the critical values for our test, we also propose a version of multiplier bootstrap and establish its asymptotic validity. Simulations show that our test behaves well in finite samples. We apply our method to study Engel curves for several major nondurable expenditures in China by using a panel dataset from China Family Panel Studies (CFPS).

The paper can be extended in various directions. First, our panel data model is of nonparametric nature in the presence of IFEs and we have a single nonparametric object of interest. It is also interesting to consider more general nonparametric panel data models with more than one nonparametric project (e.g., additive models) or semiparametric panel data models with both nonparametric and parametric components that are of interest. Second, it remains unclear how to determine the number of factors in our framework. Difficulty arises because one cannot apply existing methods (e.g., Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013), Su, Miao and Jin (2019)) that are developed under the large N and large T setup to our framework with large N and fixed T. Further complication is due to the partial identification nature of nonparametric panel. Third, it is possible to extend the current theoretical framework to conditional moment inequality models through the introduction of some slackness parameter. This will greatly broadens the scope of the current paper. We leave the extensions for future research.

APPENDIX

The appendix contains the proofs of the main results in the paper. In proving these results, we make use of several lemmas whose proofs can be found in the online supplement.

A Proofs of the main results

Let $\mathrm{MDD}(\varepsilon|W) = \left[\mathrm{MDD}(\varepsilon|W)^2\right]^{1/2}$. To prove the main results, we make use of the following lemmas.

Lemma A.1 *Let Z be a real random vector s.t. $\mathbb{E}|Z| < \infty$. For any real-valued random variables W_1 and W_2, if $\mathrm{MDD}(W_1|Z)^2 = 0$ a.s., then $\mathrm{MDD}(W_2 - W_1|Z)^2 = \mathrm{MDD}(W_2|Z)^2$ and $\mathbb{E}[(W_2 - W_1)(W_2^\dagger - W_1^\dagger) \times |Z - Z^\dagger|] = \mathbb{E}\left[W_2 W_2^\dagger |Z - Z^\dagger|\right]$, where $W_1^\dagger, W_2^\dagger$, and $Z^\dagger$ are independent copies of W_1, W_2, and Z, respectively.*

Lemma A.2 *Let Z be a real random vector s.t. $\mathbb{E}|Z| < \infty$. Let $\mathcal{W}$ be a set of real-valued random variables with uniformly bounded second moment, i.e., $\sup_{W\in\mathcal{W}} \mathbb{E}\left(W^2\right) < \infty$. Then there exists a finite constant b, s.t. for any $W_1, W_2 \in \mathcal{W}$, it holds that*

$$\left|\mathrm{MDD}(W_1|Z)^2 - \mathrm{MDD}(W_2|Z)^2\right| \leq b\,\mathrm{MDD}(W_1 - W_2|Z) \leq 2b\left[\mathrm{MDD}(W_1|Z) + \mathrm{MDD}(W_2|Z)\right].$$

Lemma A.3 *The parameter space Θ is compact under the norm $\|\cdot\|_c$ as defined by (2.9). Consequently, there exists a constant $B_c < \infty$ s.t. for all $g \in \mathcal{G}$, $\sup_{x\in\mathcal{X}} \left|D^\lambda g(x)\right| \leq B_c$ for any vector of nonnegative integers λ with $\langle\lambda\rangle \leq \frac{d}{2}$. In particular, for all $g \in \mathcal{G}$, $\sup_{x\in\mathcal{X}} |g(x)| \leq B_c$.*

Lemma A.4 *Let Assumption 2.1 and 3.1(i) hold. Define $\mathcal{Q}(\theta) \equiv \sum_{s=1}^{T-R} \mathrm{MDD}\left[m_s(Y, X, \theta)\,|\underline{z}_s\right]^2$. Then $\mathcal{Q}(\cdot)$ is Lipschitz continuous w.r.t. $\|\cdot\|_{L^2}$ in Θ.*

Lemma A.5 *Consider a generic econometric model $Q(\theta) = 0$, the identified set of which is characterized by $\Theta_I \equiv \{\theta \in \Theta : Q(\theta) = 0 \text{ a.s.}\}$. Suppose the following conditions hold: (i) $Q(\cdot) \geq 0$ and Θ is compact under (pseudo-)metric $\mathrm{d}(\cdot,\cdot)$; (ii) $\Theta_N \subseteq \Theta$ are closed and s.t. $\exists\, \Pi_N\theta$ for each $\theta \in \Theta$ s.t. $\mathrm{d}(\Pi_N\theta, \theta) = o(1)$ and $\sigma_N \equiv \sup_{\theta^0\in\Theta_I} \mathrm{d}\left(\Pi_N\theta^0, \theta^0\right) = o(1)$; (iii) $\sup_{\theta\in\Theta_N} |Q_N(\theta) - Q(\theta)| = O_p(b_N)$ for some $b_N = o(1)$; (iv) $\exists$ positive constants a_1 and a_2 s.t. $a_1\,\mathrm{d}(\theta, \Theta_I)^2 \leq Q(\theta) \leq a_2\,\mathrm{d}(\theta, \Theta_I)^2$. Then for $\hat{\theta}_N \in \operatorname{argmin}_{\theta\in\Theta_N} Q_N(\theta)$, it holds that $\mathrm{d}(\hat{\theta}_N, \Theta_I) = O_p\left(\max\{\sigma_N, b_N^{1/2}\}\right)$.*

Lemma A.6 *(i) Let Assumptions 2.1, 2.2, 3.1, 3.2, and 3.3(i) hold. For any $\hat{\theta}_N \in \operatorname{argmin}_{\theta\in\Theta_N\cap\Theta_R} S_N(\theta)$, it holds that $\mathrm{d}_{\|\cdot\|_{L^2}}\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) = o_p(1)$; (ii) If, in addition, Assumption 3.3(ii), (iii), and 3.4 hold, then it holds that $\mathrm{d}_{\|\cdot\|_{L^2}}\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) = O_p\left(\varrho_N\,\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) + \delta_{s,N}\right)$.*

Proof of Lemma 2.1. Recall from (2.10) that $\Theta_I = \{\theta = (\phi', g)' \in \boldsymbol{\Phi} \times \mathcal{G} : \mathbb{E}\left[m_s(Y_i, \phi_s, \mathbf{g}(X_i))\,|\underline{z}_{is}\right] = 0$ a.s. for $s = 1, ..., T - R\}$. Let

$$\tilde{\Theta}_I \equiv \left\{\theta = (\phi', g)' \in \Theta : \begin{array}{l} \text{For some } R\text{-dimensional random vector } \lambda_i, \text{ it holds} \\ \mathbb{E}\left[y_{it} - g(x_{it}) - \lambda_i'\phi_t|\underline{z}_{it}\right] = 0 \text{ a.s. for } t = 1, ..., T - R \\ \mathbb{E}\left[y_{it} - g(x_{it}) - \lambda_i'\left(-\iota_{t-(T-R)}\right)|\underline{z}_{it}\right] = 0 \text{ a.s. for } t = T - R + 1, ..., T \end{array}\right\} \quad \text{(A.1)}$$

where ι_r is the r'th column of the $R \times R$ identity matrix. We need to show that $\Theta_I = \tilde{\Theta}_I$.

For any given $\tilde{\theta} = (\tilde{\phi}', \tilde{g})' \in \tilde{\Theta}_I$, it holds that

$$\begin{pmatrix} \mathbb{E}\left[y_{i1} - \tilde{g}(x_{i1}) - \lambda_i' \tilde{\phi}_1 \; | \underline{z}_{i1}\right] \\ \vdots \\ \mathbb{E}\left[y_{i,T-R} - \tilde{g}(x_{i,T-R}) - \lambda_i' \tilde{\phi}_{T-R} \; | \underline{z}_{i,T-R}\right] \\ \mathbb{E}\left[y_{i,T-R+1} - \tilde{g}(x_{i,T-R+1}) - \lambda_i' \left(-\iota_1\right) \; | \underline{z}_{i,T-R+1}\right] \\ \vdots \\ \mathbb{E}\left[y_{iT} - \tilde{g}(x_{iT}) - \lambda_i' \left(-\iota_R\right) \; | \underline{z}_{iT}\right] \end{pmatrix} = \mathbf{0} \text{ a.s.} \tag{A.2}$$

By (2.6), multiplying both sides of (A.2) by the $(T-R) \times T$ matrix $H(\tilde{\phi})' \equiv (I_{T-R}, \tilde{\Phi})$ yields

$$\begin{pmatrix} \mathbb{E}\left[m_1\left(Y_i, \tilde{\phi}_1, \tilde{\mathbf{g}}(X_i)\right) | \underline{z}_{i1}\right] \\ \vdots \\ \mathbb{E}\left[m_{T-R}\left(Y_i, \tilde{\phi}_{T-R}, \tilde{\mathbf{g}}(X_i)\right) | \underline{z}_{i,T-R}\right] \end{pmatrix} = \mathbf{0} \text{ a.s.,}$$

which clearly implies that $\tilde{\theta} \in \Theta_I$. Since this holds for any $\tilde{\theta} \in \tilde{\Theta}_I$, it holds that $\tilde{\Theta}_I \subseteq \Theta_I$.

Next, for any given $\theta^0 = \left(\phi^{0\prime}, g^0\right)' \in \Theta_I$, it holds that

$$\begin{pmatrix} \mathbb{E}\left[m_1\left(Y_i, \phi_1^0, \mathbf{g}^0(X_i)\right) | \underline{z}_{i1}\right] \\ \vdots \\ \mathbb{E}\left[m_{T-R}\left(Y_i, \phi_{T-R}^0, \mathbf{g}^0(X_i)\right) | \underline{z}_{i,T-R}\right] \end{pmatrix} = \mathbf{0} \text{ a.s.,}$$

or, equivalently,

$$\begin{pmatrix} \mathbb{E}\left\{y_{i1} - g^0\left(x_{i1}\right) + \sum_{r=1}^R \phi_{1r}^0 \left[y_{i,T-R+r} - g^0\left(x_{i,T-R+r}\right)\right] | \underline{z}_{i1}\right\} \\ \vdots \\ \mathbb{E}\left\{y_{i,T-R} - g^0\left(x_{i,T-R}\right) + \sum_{r=1}^R \phi_{T-R,r}^0 \left[y_{i,T-R+r} - g^0\left(x_{i,T-R+r}\right)\right] | \underline{z}_{i1}\right\} \end{pmatrix} = \mathbf{0} \text{ a.s.} \tag{A.3}$$

Let $\lambda_i^0 \equiv (y_{i,T-R+1} - g^0\left(x_{i,T-R+1}\right), ..., y_{i,T} - g^0\left(x_{i,T}\right))'$. Then by (A.3) and the fact that $\lambda_i^{0\prime}\left(-\iota_r\right) = y_{i,T-R+r} - g^0\left(x_{i,T-R+r}\right)$, it holds that

$$\begin{pmatrix} \mathbb{E}\left[y_{i1} - g^0(x_{i1}) - \lambda_i^{0\prime} \phi_1^0 \; | \underline{z}_{i1}\right] \\ \vdots \\ \mathbb{E}\left[y_{i,T-R} - g^0(x_{i,T-R}) - \lambda_i^{0\prime} \phi_{T-R}^0 \; | \underline{z}_{i,T-R}\right] \\ \mathbb{E}\left[y_{i,T-R+1} - g^0(x_{i,T-R+1}) - \lambda_i^{0\prime} \left(-\iota_1\right) \; | \underline{z}_{i,T-R+1}\right] \\ \vdots \\ \mathbb{E}\left[y_{iT} - g^0(x_{iT}) - \lambda_i^{0\prime} \left(-\iota_R\right) \; | \underline{z}_{iT}\right] \end{pmatrix} = \mathbf{0} \text{ a.s.,} \tag{A.4}$$

which clearly implies that $\theta^0 \in \tilde{\Theta}_I$. Since it holds for any $\theta^0 \in \Theta_I$, it holds that $\Theta_I \subseteq \tilde{\Theta}_I$. This completes the proof of the lemma. ■

Proof of Lemma 3.1. Note that for any real-valued random variable W and real vector-valued random variable Z,

$$\text{MDD}\left(W|Z\right)^2 = \text{MDD}_o\left(W|Z\right)^2 + \left[\mathbb{E}\left(W\right)\right]^2 \mathbb{E}\left|Z - Z^\dagger\right|,$$

which follows directly from the definitions of MDD_o and MDD. By Definition 3.1, we have that for any $\theta_1 = (\phi_1', g_1)' \in \Theta$ and $\theta_2 = (\phi_2', g_2)' \in \Theta$,

$$\begin{aligned}
& \mathrm{d}_w(\theta_1, \theta_2)^2 \\
= & \sum_{s=1}^{T-R} \mathrm{MDD}\left[m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2) | \underline{z}_s\right]^2 \\
= & \sum_{s=1}^{T-R} \mathrm{MDD}_o\left[m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2) | \underline{z}_s\right]^2 + \sum_{s=1}^{T-R} \left\{\mathbb{E}\left[m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2)\right]\right\}^2 \mathbb{E}\left|\underline{z}_s - \underline{z}_s^{\dagger}\right|. \quad (A.5)
\end{aligned}$$

The rest of the proof is organized into three parts. In Part I, we show the existence of a constant $c_1 < \infty$ s.t.

$$\sum_{s=1}^{T-R} \mathrm{MDD}_o\left[m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2) | \underline{z}_s\right]^2 \leq c_1 \|\theta_1 - \theta_2\|_{L^2}^2.$$

In Part II, we show the existence of a constant $c_2 < \infty$ s.t.

$$\sum_{s=1}^{T-R} \left\{\mathbb{E}\left[m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2)\right]\right\}^2 \mathbb{E}\left|\underline{z}_s - \underline{z}_s^{\dagger}\right| \leq c_2 \|\theta_1 - \theta_2\|_{L^2}^2.$$

And in Part III, we combine the results from Parts I and II to complete the proof.

Part I. The compactness of $\mathbf{\Phi}$ according to Assumption 2.1(i) implies that $B_{\mathbf{\Phi}} \equiv \sup_{\phi \in \mathbf{\Phi}} |\phi| < \infty$. By Lemma A.3, for all $g \in \mathcal{G}$, it holds that $\sup_{x \in \mathcal{X}} |g(x)| \leq B_c < \infty$. Note that $m_s(Y, X, \theta) = [y_s - g(x_s)] + \sum_{r=1}^{R} \phi_{s,r}[y_{T-R+r} - g(x_{T-R+r})]$. Then for any $\theta_1 = (\phi_1', g_1)' \in \Theta$ and $\theta_2 = (\phi_2', g_2)' \in \Theta$, we have

$$\begin{aligned}
m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2) = & -[g_1(x_s) - g_2(x_s)] + \sum_{r=1}^{R} (\phi_{1,s,r} - \phi_{2,s,r})[y_{T-R+r} - g_1(x_{T-R+r})] \\
& - \sum_{r=1}^{R} \phi_{2,s,r}[g_1(x_{T-R+r}) - g_2(x_{T-R+r})]. \quad (A.6)
\end{aligned}$$

Then by the triangle inequality, we have

$$\begin{aligned}
& |m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2)| \\
\leq & |g_1(x_s) - g_2(x_s)| + \sum_{r=1}^{R} \left|\phi_{1,s,r} - \phi_{2,s,r}\right| |y_{T-R+r} - g_1(x_{T-R+r})| + \sum_{r=1}^{R} \left|\phi_{2,s,r}\right| |g_1(x_{T-R+r}) - g_2(x_{T-R+r})| \\
\leq & |g_1(x_s) - g_2(x_s)| + (|Y| + B_c) \sum_{r=1}^{R} \left|\phi_{1,s,r} - \phi_{2,s,r}\right| + B_{\mathbf{\Phi}} \sum_{r=1}^{R} |g_1(x_{T-R+r}) - g_2(x_{T-R+r})| \\
\leq & |g_1(x_s) - g_2(x_s)| + (|Y| + B_c) R |\phi_1 - \phi_2| + B_{\mathbf{\Phi}} \sum_{r=1}^{R} |g_1(x_{T-R+r}) - g_2(x_{T-R+r})|.
\end{aligned}$$

It follows that for $s = 1, ..., T-R$,

$$\begin{aligned}
& [m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2)]^2 \\
\leq & 3\left\{[g_1(x_s) - g_2(x_s)]^2 + (|Y| + B_c)^2 R^2 |\phi_1 - \phi_2|^2 + R B_{\mathbf{\Phi}}^2 \sum_{r=1}^{R} [g_1(x_{T-R+r}) - g_2(x_{T-R+r})]^2\right\} \\
\leq & B_{1m}\left\{[g_1(x_s) - g_2(x_s)]^2 + \sum_{r=1}^{R} [g_1(x_{T-R+r}) - g_2(x_{T-R+r})]^2\right\} + 3(|Y| + B_c)^2 R^2 |\phi_1 - \phi_2|^2, \quad (A.7)
\end{aligned}$$

where $B_{1m} = 3\max\left\{RB_{\Phi}^2, 1\right\}$, and the first inequality follows from the Cauchy-Schwarz inequality and Jensen inequality.

Denote by $\varphi_Z(s) \equiv \mathbb{E}\left[\exp\left(\mathbf{i}s'Z\right)\right]$ the characteristic function of Z. It holds that

$$\begin{aligned}
\left|\mathrm{Var}\left(\exp\left(\mathbf{i}s'Z\right)\right)\right| &= \left|\mathbb{E}\left[\exp\left(\mathbf{i}s'Z\right)^2\right] - \mathbb{E}\left[\exp\left(\mathbf{i}s'Z\right)\right]^2\right| = \left|\varphi_Z(2s) - \left[\varphi_Z(s)\right]^2\right| \\
&\leq \left|\varphi_Z(2s)\right| + \left|\varphi_Z(s)\right|^2 \leq 2,
\end{aligned} \tag{A.8}$$

where the last inequality follows from the fact that $|\varphi_Z(\cdot)| \leq 1$. Now, by equation (2.4) in Su and Zheng (2017),

$$\begin{aligned}
& \mathrm{MDD}_o\left[m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2)\,|\underline{z}_s\right]^2 \\
= & \int_{\mathbb{R}^{d_Z}} \left[\mathrm{Cov}\left(m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2), \exp\left(\mathbf{i}s'Z\right)\right)\right]^2 q(s)\,ds \\
\leq & \mathrm{Var}\left(m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2)\right) \int_{\mathbb{R}^{d_Z}} \left|\mathrm{Var}\left(\exp\left(\mathbf{i}s'Z\right)\right)\right| q(s)\,ds \\
\leq & 2\mathbb{E}\left\{\left[m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2)\right]^2\right\} \int_{\mathbb{R}^{d_Z}} q(s)\,ds \\
\leq & 2c_q\left[B_{1m}\mathbb{E}\left\{\left[g_1(x_s) - g_2(x_s)\right]^2 + \sum_{r=1}^{R}\left[g_1(x_{T-R+r}) - g_2(x_{T-R+r})\right]^2\right\} + 3R^2\mathbb{E}[(|Y| + B_c)^2]\left|\phi_1 - \phi_2\right|^2\right] \\
\leq & 2c_q\left[B_{1m}c_m(R+1)\left\|g_1 - g_2\right\|_{L^2}^2 + B_{2m}\left|\phi_1 - \phi_2\right|^2\right] \\
\leq & B_m\left\{\left\|g_1 - g_2\right\|_{L^2}^2 + \left|\phi_1 - \phi_2\right|^2\right\} = B_m\left\|\theta_1 - \theta_2\right\|_{L^2}^2,
\end{aligned}$$

where $\mathbf{i} \equiv \sqrt{-1}$, $q(s) \equiv 1/\left[c\,|s|^{(1+d_Z)}\right]$, $c \equiv \pi^{(1+d_Z)/2}/\Gamma\left(\frac{1+d_Z}{2}\right)$, $\Gamma(\cdot)$ is the complete gamma function: $\Gamma(z) \equiv \int_0^\infty t^{(z-1)}\exp(-t)\,dt$, $c_q = \int_{\mathbb{R}^{d_Z}} q(s)\,ds < \infty$, $B_{2m} = 3R^2\mathbb{E}[(|Y| + B_c)^2] < \infty$, $c_m < \infty$ is a constant that depends on the density of X, and $B_m = 2c_q\max\left\{B_{1m}c_m(R+1), B_{2m}\right\} < \infty$. In the derivation above, the first inequality follows from Cauchy-Schwarz inequality, the second one holds by Jensen inequality and (A.8), and the third one holds by (A.7), and the fourth one holds by the boundedness of all density functions, according to Assumption 3.1(i). Consequently, we have that for any $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$,

$$\sum_{s=1}^{T-R} \mathrm{MDD}_o\left[m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2)\,|\underline{z}_s\right]^2 \leq c_1\left\|\theta_1 - \theta_2\right\|_{L^2}^2 \tag{A.9}$$

where $c_1 \equiv (T-R)B_m < \infty$.

Part II. For any $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$, we have

$$\begin{aligned}
& \left\{\mathbb{E}\left[m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2)\right]\right\}^2 \mathbb{E}\left|\underline{z}_s - \underline{z}_s^\dagger\right| \\
\leq & \left\{\mathbb{E}\left[m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2)\right]\right\}^2 \mathbb{E}\left|Z - Z^\dagger\right| \leq \mathbb{E}\left\{\left[m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2)\right]^2\right\}\mathbb{E}\left|Z - Z^\dagger\right| \\
\leq & \left[B_{1m}\mathbb{E}\left\{\left[g_1(x_s) - g_2(x_s)\right]^2 + \sum_{r=1}^{R}\left[g_1(x_{T-R+r}) - g_2(x_{T-R+r})\right]^2\right\} + 3R^2\mathbb{E}[(|Y| + B_c)^2]\left|\phi_1 - \phi_2\right|^2\right] \\
& \times\mathbb{E}\left|Z - Z^\dagger\right| \\
\leq & \left[B_{1m}c_m(R+1)\left\|g_1 - g_2\right\|_{L^2}^2 + B_{2m}\left|\phi_1 - \phi_2\right|^2\right]\mathbb{E}\left|Z - Z^\dagger\right| \\
\leq & \tilde{B}_m\left\|\theta_1 - \theta_2\right\|_{L^2}^2,
\end{aligned}$$

where $\tilde{B}_m = \max\{B_{1m}c_m(R+1), B_{2m}\}\mathbb{E}\left|Z - Z^\dagger\right| < \infty$, the second inequality holds by Jensen inequality, and the third one follows from (A.7). Consequently, we have that for any $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$,

$$\sum_{s=1}^{T-R}\{\mathbb{E}[m_s(Y,X,\theta_1) - m_s(Y,X,\theta_2)]\}^2\mathbb{E}\left|\underline{z}_s - \underline{z}_s^\dagger\right| \leq c_2\|\theta_1 - \theta_2\|_{L^2}^2, \tag{A.10}$$

where $c_2 \equiv (T-R)\tilde{B}_m < \infty$.

Part III. Combining (A.5), (A.9) and (A.10) yields that for any $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$, $\mathrm{d}_w(\theta_1,\theta_2)^2 \leq (c_1+c_2)\|\theta_1 - \theta_2\|_{L^2}^2 = c^2\|\theta_1-\theta_2\|_{L^2}^2$, where $c \equiv \sqrt{c_1+c_2}$. This proves the first claim in Lemma 3.1.

The second claim in Lemma 3.1 follows immediately from the second inequality result of Lemma A.2. ■

To prove Theorem 3.1, we introduce some notations adopted from Arcones and Giné (1993) and de la Peña and Giné (1999, Chapter 5). Let $(S, \mathcal{S}, P)$ be a probability space and $\{\xi_i\}_{i=1}^N$ be an i.i.d. sequence with probability law P. Let $\mathcal{F}$ be a class of measurable real functions on S^m. The mth order U-process based on P and indexed by $\mathcal{F}$ is

$$U_N^m(f) \equiv U_N^m(f;P) \equiv \frac{(N-m)!}{N!}\sum_{\mathbf{i}_m \in I_N^m} f\left(\xi_{i_1},...,\xi_{i_m}\right),\ f \in \mathcal{F}, \tag{A.11}$$

where $\mathbf{i}_m \equiv (i_1,...,i_m)$, $I_N^m = \{(i_1,...,i_m) : i_j \in \mathbb{N},\ 1 \leq i_j \leq N, \text{ and } i_j \neq i_k \text{ if } j \neq k\}$. We will repeatedly use the Hoeffding's decomposition of a U-statistic. The operator $\pi_{k,m} = \pi_{k,m}^P$ acts on P^m-integrable function $f : S^m \to \mathbb{R}$ as follows

$$\pi_{k,m}f(\xi_1,...,\xi_k) = \left(\delta_{\xi_1} - P\right)\cdots\left(\delta_{\xi_k} - P\right)P^{m-k}f, \tag{A.12}$$

where δ_{ξ_j} is the Dirac measure at the observation ξ_j. Note that $\pi_{k,m}f$ is a P-canonical function of k variables. Then we have the following Hoeffding's decomposition

$$U_N^m(f) = \sum_{k=0}^m \begin{pmatrix} m \\ k \end{pmatrix} U_N^m(\pi_{k,m} \circ S_m f), \tag{A.13}$$

where $S_m f$ is a symmetric version of f : $S_m f(\xi_1,...,\xi_k) = (m!)^{-1}\sum f\left(\xi_{i_1},...,\xi_{i_m}\right)$ with the sum extended over $m!$ permutations $(i_1,...,i_m)$ of $\{1,...,m\}$.

Given a pseudometric space $(\mathcal{F},\ e)$, the ε-covering number of $(\mathcal{F},\ e)$ is

$$\mathbb{N}(\varepsilon, \mathcal{F},\ e) \equiv \min\left\{n : \exists f_1,...,f_n \in \mathcal{F} \text{ s.t. } \sup_{f\in\mathcal{F}}\min_{i\leq n} e(f,f_i) \leq \varepsilon\right\}.$$

We define $\mathbb{N}_{N,p}(\epsilon,\mathcal{F}) \equiv \mathbb{N}_{N,p}(\epsilon,\mathcal{F},e_{N,p})$ as the *random* ε-covering numbers of $(\mathcal{F}, e_{N,p})$, where $e_{N,p}(f,g) = \{U_N^m(|f-g|^p)\}^{1/p}$ where $p \geq 1$. Note that $e_{N,p}$ denotes the L^p distance corresponding to the random measure that assigns mass $\frac{(N-m)!}{N!}$ to each of the points $(\xi_{i_1},...,\xi_{i_m}) \in S^m$, $\mathbf{i}_m \in I_N^m$. When $\mathcal{F}$ is a class of real symmetric measurable functions on S^m, we define pseudo-distances $e_{N,k,2}$ on $\mathcal{F}$ as follows

$$e_{N,k,2}^2(f,g) \equiv \frac{N^k}{\binom{N}{K}}U_N^k\left(\pi_{k,m}(f-g)^2\right).$$

By the proof of Corollary 5.7 in Arcones and Giné (1993), there exist some positive finite constants $c_{k,r}$ such that for all $\varepsilon > 0$,

$$\mathbb{N}_{N,2}\left(\varepsilon, \pi_{k,m}\mathcal{F}\right) \leq \prod_{r=0}^{k} \mathbb{N}\left(\frac{\varepsilon}{2\left(k+1\right)^{1/2} c_{k,2}}, \mathcal{F},\ \left\|\cdot\right\|_{L^2\left(U_N^r \times P^{m-r}\right)}\right), \tag{A.14}$$

where for $r > 0$, $U_N^r \times P^{m-r}$ denotes the random probability measure

$$U_N^r \times P^{m-r} = \frac{(N-r)!}{N!} \sum_{\mathbf{i}_r \in I_N^r} \delta_{\left(\xi_{i_1},\ldots,\xi_{i_r}\right)} \times P^{m-r}$$

defined on $(S^m, \mathcal{S}^m)$ and for $r = 0$, $U_N^0 \times P^m$ just means P^m. Here $L^2\left(U_N^r \times P^{m-r}\right)$ defines the pseudometric on S^m :

$$\|f - g\|_{L^2\left(U_N^r \times P^{m-r}\right)}^2 = U_N^r \times P^{m-r}\left(f - g\right)^2.$$

Note that

$$\mathbb{N}\left(\varepsilon, \mathcal{F},\ \left\|\cdot\right\|_{L^2\left(U_N^0 \times P^m\right)}\right) \simeq 2P^m F/\varepsilon \text{ if } \varepsilon \leq 2P^m F \text{ and equals 1 otherwise} \tag{A.15}$$

Proof of Theorem 3.1. Note that $S_{Ns}(\theta) = S_{Ns,1}(\theta) + S_{Ns,2}(\theta)$, where $S_{Ns,1}(\theta) = -\frac{1}{N}\sum_{1\leq i\neq j\leq N} m_{is}(\theta) \times m_{js}(\theta)\kappa_{ij,s}$ and $S_{Ns,2}(\theta) = \frac{2}{N}\sum_{1\leq i\neq j\leq N} m_{is}(\theta)\kappa_{ij,s}\frac{1}{N}\sum_{k=1}^{N} m_{ks}(\theta)$. Let $\bar{m}_{is}(\theta) = \mathbb{E}\left[m_{is}(\theta)|\underline{z}_{is}\right]$ and $\tilde{m}_{is}(\theta) = m_{is}(\theta) - \bar{m}_{is}(\theta)$. Then

$$\begin{aligned}
S_{Ns,1}(\theta) &= -\frac{1}{N}\sum_{1\leq i\neq j\leq N}\left[\tilde{m}_{is}(\theta) + \bar{m}_{is}(\theta)\right]\left[\tilde{m}_{js}(\theta) + \bar{m}_{js}(\theta)\right]\kappa_{ij,s} \\
&= -\frac{1}{N}\sum_{1\leq i\neq j\leq N}\tilde{m}_{is}(\theta)\tilde{m}_{js}(\theta)\kappa_{ij,s} - \frac{1}{N}\sum_{1\leq i\neq j\leq N}\bar{m}_{is}(\theta)\bar{m}_{js}(\theta)\kappa_{ij,s} - \frac{2}{N}\sum_{1\leq i\neq j\leq N}\tilde{m}_{is}(\theta)\bar{m}_{js}(\theta)\kappa_{ij,s},
\end{aligned}$$

and

$$\begin{aligned}
S_{Ns,2}(\theta) &= \frac{2}{N^2}\sum_{1\leq i\neq j\leq N}\sum_{k=1}^{N}\tilde{m}_{is}(\theta)\tilde{m}_{ks}(\theta)\kappa_{ij,s} + \frac{2}{N^2}\sum_{1\leq i\neq j\leq N}\sum_{k=1}^{N}\bar{m}_{is}(\theta)\bar{m}_{ks}(\theta)\kappa_{ij,s} \\
&\quad + \frac{2}{N^2}\sum_{1\leq i\neq j\leq N}\sum_{k=1}^{N}\left[\tilde{m}_{is}\bar{m}_{ks}(\theta) + \bar{m}_{is}(\theta)\tilde{m}_{ks}(\theta)\right]\kappa_{ij,s}.
\end{aligned}$$

Then $S_{Ns}(\theta) = \tilde{S}_{Ns,1}(\theta) + \tilde{S}_{Ns,2}(\theta) + \tilde{S}_{Ns,3}(\theta)$, where

$$\begin{aligned}
\tilde{S}_{Ns,1}(\theta) &= -\frac{1}{N}\sum_{1\leq i\neq j\leq N}\tilde{m}_{is}(\theta)\tilde{m}_{js}(\theta)\kappa_{ij,s} + \frac{2}{N^2}\sum_{1\leq i\neq j\leq N}\sum_{k=1}^{N}\tilde{m}_{is}(\theta)\tilde{m}_{ks}(\theta)\kappa_{ij,s} \\
\tilde{S}_{Ns,2}(\theta) &= -\frac{1}{N}\sum_{1\leq i\neq j\leq N}\bar{m}_{is}(\theta)\bar{m}_{js}(\theta)\kappa_{ij,s} + \frac{2}{N^2}\sum_{1\leq i\neq j\leq N}\sum_{k=1}^{N}\bar{m}_{is}(\theta)\bar{m}_{ks}(\theta)\kappa_{ij,s} \\
\tilde{S}_{Ns,3}(\theta) &= -\frac{2}{N}\sum_{1\leq i\neq j\leq N}\tilde{m}_{is}(\theta)\bar{m}_{js}(\theta)\kappa_{ij,s} + \frac{2}{N^2}\sum_{1\leq i\neq j\leq N}\sum_{k=1}^{N}\left[\tilde{m}_{is}\bar{m}_{ks}(\theta) + \bar{m}_{is}(\theta)\tilde{m}_{ks}(\theta)\right]\kappa_{ij,s}.
\end{aligned}$$

We prove parts (i) and (ii) of the theorem in turn.

Part I. Proof of part (i).

When $\theta \in \Theta_I$, $\bar{m}_{is}(\theta) = 0$ for all $i = 1, ..., N$ and $s = 1, ..., T-R$. This implies that $\tilde{S}_{Ns,2}(\theta) = \tilde{S}_{Ns,3}(\theta) = 0$. We are left to study $\tilde{S}_{Ns,1}(\theta)$. For the second term in the definition of $\tilde{S}_{Ns,1}(\theta)$, we have

$$\begin{aligned}
& \frac{2}{N} \sum_{1 \leq i \neq j \leq N} \sum_{k=1}^{N} \tilde{m}_{is}(\theta) \tilde{m}_{ks}(\theta) \kappa_{ij,s} \\
= & \frac{2}{N^2} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta)^2 \kappa_{ij,s} + \frac{2}{N^2} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta) \tilde{m}_{js}(\theta) \kappa_{ij,s} + \frac{2}{N^2} \sum_{1 \leq i \neq j \neq k \leq N} \tilde{m}_{is}(\theta) \tilde{m}_{ks}(\theta) \kappa_{ij,s} \\
= & \frac{2}{N^2} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta)^2 \kappa_{ij,s} + \frac{2}{N^2} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta) \tilde{m}_{js}(\theta) \kappa_{ij,s} + \frac{(N-1)(N-2)}{N^2} N \mathbb{U}_{2Ns}
\end{aligned}$$

where $\mathbb{U}_{2Ns} = \binom{N}{3}^{-1} \sum_{1 \leq i < j \leq k \leq N} \psi_s(\xi_i, \xi_j, \xi_k; \theta)$ and $\psi_s(\xi_i, \xi_j, \xi_k; \theta) = \frac{1}{3}[\tilde{m}_{is}(\theta) \tilde{m}_{ks}(\theta) \kappa_{ij,s} + \tilde{m}_{is}(\theta) \times \tilde{m}_{js}(\theta) \kappa_{ik,s} + \tilde{m}_{js}(\theta) \tilde{m}_{ks}(\theta) \kappa_{jk,s} + \tilde{m}_{js}(\theta) \tilde{m}_{is}(\theta) \kappa_{jk,s} + \tilde{m}_{ks}(\theta) \tilde{m}_{is}(\theta) \kappa_{jk,s} + \tilde{m}_{ks}(\theta) \tilde{m}_{js}(\theta) \kappa_{ik,s}]$ is a symmetrized version of $\psi_{0s}(\xi_i, \xi_j, \xi_k; \theta) \equiv 2\tilde{m}_{is}(\theta) \tilde{m}_{ks}(\theta) \kappa_{ij,s}$. Note that

$$\begin{aligned}
\mathbb{E}[\psi_s(\xi_1, \xi_2, \xi_3; \theta)] &= 0, \ \mathbb{E}[\psi(\xi_1, \xi_2, \xi_3; \theta) | \xi_1] = 0 \text{ and} \\
\mathbb{E}[\psi_s(\xi_1, \xi_2, \xi_3; \theta) | \xi_1, \xi_2] &= \frac{1}{3} \tilde{m}_{1s}(\theta) \tilde{m}_{2s}(\theta) [\mathbb{E}_3(\kappa_{13,s}) + \mathbb{E}_3(\kappa_{23,s})] \equiv h_s^{(2)}(\xi_1, \xi_2; \theta)
\end{aligned}$$

Let $h_s^{(3)}(\xi_1, \xi_2, \xi_3; \theta) = \psi_s(\xi_1, \xi_2, \xi_3; \theta) - [h_s^{(2)}(\xi_1, \xi_2; \theta) + h_s^{(2)}(\xi_1, \xi_3; \theta) + h_s^{(2)}(\xi_2, \xi_3; \theta)]$. By Hoeffding's decomposition in (A.13) (see also Lee (1996, p.26) and de la Peña and Giné (1999, p.137)), we have $\mathbb{U}_{2Ns}(\theta) = 3\mathbb{H}_{2Ns}(\theta) + \mathbb{H}_{3Ns}(\theta)$, where

$$\mathbb{H}_{2Ns}(\theta) = \binom{N}{2}^{-1} \sum_{1 \leq i < j \leq N} h_s^{(2)}(\xi_i, \xi_j; \theta) \text{ and } \mathbb{H}_{3Ns}(\theta) = \binom{N}{3}^{-1} \sum_{1 \leq i < j \leq k \leq N} h_s^{(3)}(\xi_i, \xi_j, \xi_k; \theta).$$

Similarly, we can write the first term in the definition of $\tilde{S}_{Ns,1}(\theta)$ as follows: $-\frac{1}{N} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta) \tilde{m}_{js}(\theta) \kappa_{ij,s}$ $= \frac{N-1}{N} N \mathbb{U}_{1Ns}(\theta)$, where $\mathbb{U}_{1N}(\theta) = -\left(\begin{array}{c} N \\ 2 \end{array}\right)^{-1} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta) \tilde{m}_{js}(\theta) \kappa_{ij,s}$. Then we have

$$\begin{aligned}
\tilde{S}_{Ns,1}(\theta) &= \frac{N-1}{N} N \mathbb{U}_{1Ns}(\theta) + \frac{(N-1)(N-2)}{N^2} N \mathbb{U}_{2Ns} + \frac{2}{N^2} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta)^2 \kappa_{ij,s} \\
& \quad + \frac{2}{N^2} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta) \tilde{m}_{js}(\theta) \kappa_{ij,s} \\
&= N \mathbb{U}_{Ns}(\theta) + N \mathbb{H}_{3Ns}(\theta) + \frac{2}{N^2} \sum_{1 \leq i \neq j \leq N} \tilde{m}_{is}(\theta)^2 \kappa_{ij,s} - \frac{3N-2}{N} \mathbb{U}_{1Ns}(\theta) - \frac{3N-2}{N} \mathbb{U}_{2Ns},
\end{aligned}$$

where $\mathbb{U}_{Ns}(\theta) = \mathbb{U}_{1Ns}(\theta) + 3\mathbb{H}_{2Ns}(\theta) = \binom{N}{2}^{-1} \sum_{1 \leq i < j \leq N} h_s(\xi_i, \xi_j; \theta)$ and $h_s(\xi_i, \xi_j; \theta) = \tilde{m}_{is}(\theta) \tilde{m}_{js}(\theta) \times [\mathbb{E}_j(\kappa_{ij,s}) + \mathbb{E}_i(\kappa_{ij,s}) - \kappa_{ij,s}]$.

Let $\mathcal{X}$ and $\mathcal{Z}_s$ denote the supports of x_{it} and $\underline{z}_{is}$, respectively. Note that $m_s(Y, X; \theta) = H_s(\phi_s)' [Y - \mathbf{g}(X)]$ $= [y_s - g(x_s)] + \sum_{r=1}^{R} \phi_{s,r} [y_{T-R+r} - g(x_{T-R+r})]$ and $\phi_s = (\phi_{s,1}, ..., \phi_{s,R})'$ for $s = 1, ..., T-R$. Let $\theta_s = (\phi_s', g)'$. Define

$$\begin{aligned}
\mathcal{F}_{1s} \equiv \{m_s(\cdot, \cdot; \theta_s) : \mathbb{R}^T \times \mathcal{X}^T \to \mathbb{R} : \ m_s(y, x; \theta_s) = [y_s - g(x_s)] + \sum_{r=1}^{R} \phi_{s,r} [y_{T-R+r} - g(x_{T-R+r})] \\
\text{for some } \theta_s = (\phi_s', g)' \in \Phi_s \times \mathcal{G}\}, \qquad \text{(A.16)}
\end{aligned}$$

where $\Phi_s \equiv \left\{\phi_s \in \mathbb{R}^R : \|\phi_s\| \leq c_\phi\right\}$ for some constant c_ϕ, and $\mathcal{G} \equiv \{g \in W^s(\mathcal{X}) : \|g\|_s \leq B\}$. Similarly, let $\xi = (y', x', z')'$ and $S = \mathbb{R}^T \times \mathcal{X}^T \times \mathcal{Z}^T$. Define

$$\begin{aligned}
\mathcal{F}_{1s}^c \equiv \{\tilde{m}_s(\cdot;\theta_s) : S \to \mathbb{R} :\ & \tilde{m}_s(\xi;\theta_s) = [y_s - g(x_s)] - \mathbb{E}[(y_{is} - g(x_{is}))|\underline{z}_{is} = \underline{z}_s] \\
+ \sum_{r=1}^R \phi_{s,r}\{[y_{T-R+r} - g(x_{T-R+r})] & - \mathbb{E}[(y_{i,T-R+r} - g(x_{i,T-R+r}))|\underline{z}_{is} = \underline{z}_s]\} \\
& \text{for some } \theta_s = (\phi_s', g)' \in \Phi_s \times \mathcal{G}\},
\end{aligned} \tag{A.17}$$

$$\begin{aligned}
\mathcal{F}_{2s} \equiv \{f_s(\cdot,\cdot;\theta_s) : S \times S \to \mathbb{R} :\ & f_s(\xi_1,\xi_2;\theta_s) = \tilde{m}_s(\xi_1;\theta_s)\tilde{m}_s(\xi_2;\theta_s)\breve{\kappa}_{12,s} \\
& \text{for some } \theta_s = (\phi_s', g)' \in \Phi_s \times \mathcal{G}\},
\end{aligned} \tag{A.18}$$

and

$$\begin{aligned}
\mathcal{F}_{3s} \equiv \{f_s(\cdot,\cdot,\cdot;\theta_s) : S \times S \times S \to \mathbb{R} :\ & f_s(\xi_1,\xi_2,\xi_3;\theta_s) = \tilde{m}_s(\xi_1;\theta_s)\tilde{m}_s(\xi_2;\theta_s)\breve{\kappa}_{12,s} \\
+\tilde{m}_s(\xi_1;\theta_s)\tilde{m}_s(\xi_3;\theta_s)\breve{\kappa}_{13,s} & + \tilde{m}_s(\xi_2;\theta_s)\tilde{m}_s(\xi_3;\theta_s)\breve{\kappa}_{23,s} \\
& \text{for some } \theta_s = (\phi_s', g)' \in \Phi_s \times \mathcal{G}\},
\end{aligned} \tag{A.19}$$

where, e.g., $\underline{z}_{is} = (z_{i1}', ..., z_{is}')'$ and $\breve{\kappa}_{ij,s} = \mathbb{E}_j(\kappa_{ij,s}) + \mathbb{E}_i(\kappa_{ij,s}) - \kappa_{ij,s}$. The compactness of $\boldsymbol{\Phi}$ according to Assumption 2.1(i) implies that $B_{\boldsymbol{\Phi}} \equiv \sup_{\phi \in \boldsymbol{\Phi}} |\phi| < \infty$. By Lemma A.3, for all $g \in \mathcal{G}$, it holds that $\sup_{x \in \mathcal{X}} |g(x)| \leq B_c < \infty$. Then for any $\theta \in \Theta$ and $s = 1, ..., T-R$, we have

$$\begin{aligned}
|m_s(Y, X, \theta)| &\leq [|y_s| + |g(x_s)|] + \sum_{r=1}^R |\phi_{s,r}| [|y_{T-R+r}| + |g(x_{T-R+r})|] \\
&\leq |Y| + B_c + B_{\boldsymbol{\Phi}} R [|Y| + B_c] = (B_{\boldsymbol{\Phi}} R + 1)[|Y| + B_c] \leq K[|Y| + 1] \equiv F_1(Y),
\end{aligned} \tag{A.20}$$

where $Y = (y_1, ..., y_T)'$, $X = (x_1, ..., x_T)'$, $\phi_{s,r}$ denote the rth element in ϕ_s, and K is a generic positive constant that may vary across lines. By (A.6),

$$\begin{aligned}
|m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2)| &\leq (|Y| + B_c)R|\phi_1 - \phi_2| + (RB_{\boldsymbol{\Phi}} + 1)\|g_1 - g_2\|_\infty \\
&\leq K(|Y| + 1)\{|\phi_1 - \phi_2| + \|g_1 - g_2\|_\infty\}.
\end{aligned}$$

It follows that the class $\mathcal{F}_{1s}$ is Lipschitz in $\Phi_s \times \mathcal{G}$ (w.r.t.) the norm $|\cdot| + \|\cdot\|_\infty$. Then by Theorem 2.7.11 in van der Vaart and Wellner (1996), we have

$$\begin{aligned}
\mathbb{N}_{[]}(\epsilon\|F_1\|,\ \mathcal{F}_{1s},\ \|\cdot\|_{L^2}) &\leq \mathbb{N}(\epsilon/2,\ \Phi_s \times \mathcal{G},\ |\cdot| + \|\cdot\|_\infty) \leq \mathbb{N}(\epsilon/4,\ \Phi_s,\ |\cdot|)\,\mathbb{N}(\epsilon/4,\ \mathcal{G},\ \|\cdot\|_\infty) \\
&\leq K\left(\frac{4}{\epsilon}\right)^R \exp\left[\left(\frac{4}{\epsilon}\right)^\nu\right],
\end{aligned} \tag{A.21}$$

where the first inequality follows from Theorem 2.7.11 in van der Vaart and Wellner (1996) and the last one follows from Lemma A.3 in Santos (2012), which indicates that $\mathbb{N}(\epsilon, \mathcal{G}, \|\cdot\|_\infty) \leq K\exp\left(\left(\frac{1}{\epsilon}\right)^\nu\right)$ with ν being defined in (2.8). This also implies that

$$\mathbb{N}_{[]}(2\epsilon\|F\|,\ \mathcal{F}_{1s}^c,\ \|\cdot\|_{L^2}) \leq K\left(\frac{4}{\epsilon}\right)^{2R} \exp\left[2\left(\frac{4}{\epsilon}\right)^\nu\right]. \tag{A.22}$$

Let $F_2(\xi_1, \xi_2) = K(|y_1|+1)(|y_2|+1)(|z_1|+|z_2|+1)$ with $\xi_i = (y_i', x_i', z_i')' \in S$. By arguments as used in the proof of Theorem 6 in Andrews (1994), there is a finite positive constant c_0 such that

$$\mathbb{N}_{[]}\left(2c_0\epsilon \|F_2\|,\ \mathcal{F}_{2s},\ \|\cdot\|\right) \le \left[\mathbb{N}_{[]}\left(2\epsilon\|F_1\|,\ \mathcal{F}_{2s},\ \|\cdot\|\right)\right]^2 \le K\left(\frac{4}{\epsilon}\right)^{4R} \exp\left[4\left(\frac{4}{\epsilon}\right)^{\nu}\right]. \tag{A.23}$$

We verify the conditions in Theorem 5.6 of Arcones and Giné (1993, AC hereafter). First, by Assumption 3.1(i)-(ii),

$$\begin{aligned}\mathbb{E}\left(\left[F_2(\xi_1,\xi_2)\right]^2\right) &\le 2K^2 E\left\{\left[(|Y_1|+1)(|Y_2|+1)\,|Z_1|\right]^2 + \left[(|Y_1|+1)(|Y_2|+1)\,|Z_2|\right]^2\right\} \\ &\le 4K^2\mathbb{E}\left[(|Y_1|+1)^2\,|Z_1|^2\right]\mathbb{E}\left[(|Y_2|+1)^2\right] \le \infty.\end{aligned}$$

This verifies Condition (a) in Theorem 5.6 of AC. Applying (A.14) with $m = k = 2$ yields

$$\mathbb{N}_{N,2}(\varepsilon, \mathcal{F}_{2s}) = \mathbb{N}_{N,2}(\varepsilon, \pi_{2,2}\mathcal{F}_{2s}) \le \prod_{r=0}^{2}\mathbb{N}\left(\frac{\varepsilon}{2\sqrt{3}c_{2,r}}, \mathcal{F}_{2s},\ \|\cdot\|_{L^2\left(U_N^r \times P^{2-r}\right)}\right).$$

By (A.15), $\int_0^\delta \log\mathbb{N}\left(\frac{\varepsilon}{2\sqrt{3}c_{2,0}}, \mathcal{F}_{2s},\ \|\cdot\|_{L^2\left(U_N^0\times P^2\right)}\right) d\epsilon \preceq \int_0^\delta \log\frac{1}{\varepsilon} d\epsilon$. Note that

$$\begin{aligned}&\int_0^\delta \log\mathbb{N}\left(\frac{\varepsilon}{2\sqrt{3}c_{2,1}}, \mathcal{F}_{2s},\ \|\cdot\|_{L^2\left(U_N^1\times P^1\right)}\right) d\varepsilon \\ =\ & 2\sqrt{3}c_{2,1}\left[U_N^1(P^1F_2^2)\right]^{1/2}\int_0^{\delta/\left[2\sqrt{3}c_{2,1}U_N^1(P^1F^2)\right]^{1/2}} \log N\left(\varepsilon\left[U_N^1(P^1\bar{F}^2)\right]^{1/2}, \mathcal{F}_{2s},\ \|\cdot\|_{L^2(U_N^1\times P^1)}\right) d\varepsilon \\ \preceq\ & \left[U_N^1(P^1F_2^2)\right]^{1/2}\int_0^{\delta/\left[2\sqrt{3}c_{2,1}U_N^1(P^1F_2^2)\right]^{1/2}} \left[\log\left(\frac{4}{\varepsilon}\right) + \left(\frac{4}{\varepsilon}\right)^{\nu}\right] d\varepsilon \\ \preceq\ & \left[U_N^1(P^1F_2^2)\right]^{1/2}\int_0^{\delta/\left[2\sqrt{3}c_{2,1}U_N^1(P^1F_2^2)\right]^{1/2}} \varepsilon^{-\nu} d\varepsilon \\ \preceq\ & \left[U_N^1(P^1F_2^2)\right]^{\nu/2}\delta^{1-\nu},\end{aligned}$$

where the first equality follows from the change of variables, the first inequality holds by (A.23), and the second inequality follows from the fact that the integrand is dominated by the term $(\varepsilon/4)^{-\nu}$ in the neighborhood of 0, $\nu < 1$ by Assumption 2.1(iii), and $\int_0^{\delta'} \log(1/\varepsilon)\, d\varepsilon < \infty$ for any $\delta' < \infty$. Similarly, we have

$$\begin{aligned}&\int_0^\delta \log\mathbb{N}\left(\frac{\varepsilon}{2\sqrt{3}c_{2,2}}, \mathcal{F}_{2s},\ \|\cdot\|_{L^2\left(U_N^2\right)}\right) d\varepsilon \\ =\ & 2\sqrt{3}c_{2,2}\left[U_N^2(\bar{F}^2)\right]^{1/2}\int_0^{\delta/\left[2\sqrt{3}c_{2,2}U_N^2(F^2)\right]^{1/2}} \log\mathbb{N}\left(\varepsilon\left[U_N^2(F^2)\right]^{1/2}, \mathcal{F},\ \|\cdot\|_{L^2(U_N^2)}\right) d\varepsilon \\ \preceq\ & \left[U_N^2(P^2F_2^2)\right]^{1/2}\int_0^{\delta/\left[2\sqrt{3}c_{2,2}U_N^2(F^2)\right]^{1/2}} \left[\log\left(\frac{4}{\varepsilon}\right) + \left(\frac{4}{\varepsilon}\right)^{\nu}\right] d\varepsilon \\ \preceq\ & \left[U_N^2(F_2^2)\right]^{1/2}\int_0^{\delta/\left[2\sqrt{3}c_{2,2}U_N^2(F^2)\right]^{1/2}} \varepsilon^{-\nu} d\varepsilon \preceq \left[U_N^2(F_2^2)\right]^{\nu/2}\delta^{1-\nu}.\end{aligned}$$

Then

$$\begin{aligned}
& \lim_{\delta\to 0}\limsup_{N\to\infty}\mathbb{E}^{o}\left[\int_0^{\delta}\log \mathbb{N}_{N,2}\left(\varepsilon,\mathcal{F}_{2s}\right)d\epsilon\right] \\
= & \lim_{\delta\to 0}\limsup_{N\to\infty}\mathbb{E}^{*}\left[\int_0^{\delta}\sum_{r=0}^{2}\log \mathbb{N}\left(\frac{\varepsilon}{2\sqrt{3}c_{2,r}},\mathcal{F}_{2s},\ \left\|\cdot\right\|_{L^2\left(U_N^r\times P^{2-r}\right)}\right)d\varepsilon\right] \\
\preceq & \lim_{\delta\to 0}\limsup_{N\to\infty}\left[\int_0^{\delta}\log\frac{1}{\varepsilon}d\epsilon+\mathbb{E}\left\{\left[U_N^1(P^1F_2^2)\right]^{\nu/2}+\left[U_N^2(F_2^2)\right]^{\nu/2}\right\}\delta^{1-\nu}\right]=0,
\end{aligned}$$

where the last equality follows from the fact that $\mathbb{E}\{\left[U_N^2(F_2^2)\right]^{\nu/2}\}\leq\left\{\mathbb{E}\left[U_N^2(F_2^2)\right]\right\}^{\nu/2}=\{\mathbb{E}([F_2\left(\xi_1,\xi_2\right)]^2\}^{\nu/2}<\infty$ by Jensen inequality and similarly $\mathbb{E}\{\left[U_N^1(P^1F_2^2)\right]^{\nu/2}\}<\infty$. This verifies condition (c) in Theorem 5.6 of AC. Next, notice that $\mathbb{N}\left(\frac{\varepsilon}{2\sqrt{3}c_{2,r}},\mathcal{F}_{2s},\ \left\|\cdot\right\|_{L^2\left(U_N^r\times P^{2-r}\right)}\right)=1$ a.s. for $r=0,1,2$ and for sufficiently large ε, say, $\varepsilon\geq\varepsilon_0$, by the total boundedness of $\Phi\times\mathcal{G}$ and the law of large numbers for U-statistics. It follows that for some small $\epsilon>0$ and by the above calculations,

$$\begin{aligned}
& \mathbb{E}^{o}\left|\int_0^{\infty}\log \mathbb{N}_{N,2}\left(\varepsilon,\pi_{2,2}\mathcal{F}_{2s}\right)d\varepsilon\right|^{1+\epsilon} \\
\preceq & \mathbb{E}^{o}\left|\int_0^{\varepsilon_0}\log \mathbb{N}_{N,2}\left(\varepsilon,\pi_{2,2}\mathcal{F}_{2s}\right)d\varepsilon\right|^{1+\epsilon} \\
\preceq & \left|\int_0^{\varepsilon_0}\log\frac{1}{\varepsilon}d\epsilon\right|^{1+\epsilon}+\mathbb{E}\left\{\left[U_N^1(P^1F_2^2)\right]^{(1+\epsilon)\nu/2}+\left[U_N^2(F_2^2)\right]^{(1+\epsilon)\nu/2}\right\}<\infty,
\end{aligned}$$

where $\mathbb{E}^{o}$ denotes the outer-expectation associated to $\mathbb{E}$, the last inequality holds by choosing ϵ sufficiently small such that $(1+\epsilon)\nu/2\leq 1$. This implies that the sequence $\left\{\int_0^{\infty}\log N_{N,2}\left(\varepsilon,\mathcal{F}_{2s}\right)d\varepsilon\right\}_{N=1}^{\infty}$ is uniformly integrable. That is, condition (b) in Theorem 5.6 of AC is verified. Then by Theorem 5.6 of AC, we have $N\mathbb{U}_N\left(\theta\right)\Longrightarrow\mathbb{C}_s\left(\theta\right)$ in $L^{\infty}\left(\Theta\right).$

Next, note that $\mathbb{H}_{3N}\left(\theta\right)$ is a third order $P-$canonical $U-$process with the envelope function for the kernel in the definition of $\mathbb{H}_{3N}\left(\theta\right)$ given by $F_3\left(\xi_1,\xi_2,\xi_3\right)=K\{(|y_1|+1)(|y_2|+1)(|z_1|+|z_2|+1)+(|y_1|+1)(|y_3|+1)(|z_1|+|z_3|+1)+(|y_2|+1)(|y_3|+1)(|z_2|+|z_3|+1)\}$. Following the analysis of $\mathbb{U}_N\left(\theta\right)$, it is easy to show that $\mathbb{E}[F_3\left(\xi_1,\xi_2,\xi_3\right)^2]<\infty$ under Assumption 3.1(i)-(ii),

$$\lim_{\delta\to 0}\limsup_{N\to\infty}\mathbb{E}^{o}\left[\int_0^{\delta}\left[\log \mathbb{N}_{N,2}\left(\varepsilon,\mathcal{F}_{3s}\right)\right]^{3/2}d\epsilon\right]=0$$

and the sequence $\left\{\int_0^{\infty}\left[\log N_{N,2}\left(\varepsilon,\mathcal{F}_{3s}\right)\right]^{3/2}d\varepsilon\right\}_{N=1}^{\infty}$ is uniformly integrable. Here we use the fact that $\frac{3}{2}\left(1+\epsilon\right)\nu/2\leq 1$ for sufficiently small ϵ. As a result, $N^{3/2}\mathbb{H}_{3N}\left(\theta\right)$ converges to a Gaussian chaos process and $\sup_{\theta\in\Theta}\left|N\mathbb{H}_{3N}\left(\theta\right)\right|=O_P\left(N^{-1/2}\right)$. Our condition is sufficient to ensure the uniform law of large numbers to hold for the U-process with kernel function associated with $\tilde{m}_{is}\left(\theta\right)^2\kappa_{ij,s}$. As a result, we have

$$\frac{2}{N^2}\sum_{1\leq i\neq j\leq N}\tilde{m}_{is}\left(\theta\right)^2\kappa_{ij,s}=2\mathbb{E}\left[\tilde{m}_{1s}\left(\theta\right)^2\kappa_{12,s}\right]+o_p(1)\equiv\mathbb{B}_s\left(\theta\right)+o_p(1)\text{ uniformly in }\theta\in\Theta.$$

Following the analysis of $\mathbb{U}_{Ns}\left(\theta\right)$, we can also show that both $N\mathbb{U}_{1Ns}\left(\theta\right)$ and $N\mathbb{U}_{2Ns}\left(\theta\right)$ converge to Gaussian chaos processes. Consequently, we have

$$\tilde{S}_{Ns,1}\left(\theta\right)\Longrightarrow\mathbb{B}_s\left(\theta\right)+\mathbb{C}_s\left(\theta\right). \tag{A.24}$$

When $\theta \in \Theta_I$, we also have $S_{Ns}(\theta) \Longrightarrow \mathbb{B}_s(\theta) + \mathbb{C}_s(\theta)$ given the fact that $\tilde{S}_{Ns,\ell}(\theta) = 0$ for $\ell = 2, 3$ in this case. As a result, we have $S_N(\theta) \Longrightarrow \mathbb{B}(\theta) + \mathbb{C}(\theta)$.

Part II. Proof of part (ii).

When $\theta \notin \Theta_I$, (A.24) continues to hold. By the law of large numbers for U-processes and the definition of MDD, we have

$$\begin{aligned}\frac{1}{N}\tilde{S}_{Ns,2}(\theta) &= -\frac{1}{N^2}\sum_{1\le i\ne j\le N}\bar{m}_{is}(\theta)\bar{m}_{js}(\theta)\kappa_{ij,s} + \frac{2}{N^3}\sum_{1\le i\ne j\le N}\sum_{k=1}^{N}\bar{m}_{is}(\theta)\bar{m}_{ks}(\theta)\kappa_{ij,s} \\ &= -\mathbb{E}[\bar{m}_{1s}(\theta)\bar{m}_{2s}(\theta)\kappa_{12,s}] + 2\mathbb{E}[\bar{m}_{1s}(\theta)\kappa_{12,s}]\mathbb{E}[\bar{m}_{2s}(\theta)] + o_P(1) \\ &= \mathrm{MDD}[m_s(Y,X,\theta)|\underline{z}_s]^2 + o_P(1) \text{ uniformly in } \theta \in \Theta\backslash\Theta_I.\end{aligned}$$

In fact, applying Hoeffding decomposition to $\frac{1}{N}\tilde{S}_{Ns,2}(\theta)$ and arguments as used in Part I, we can strengthen $o_P(1)$ to $O_P\left(N^{-1/2}\right)$ in the last claim.

Now, define $\mathbb{U}_{3N}(\theta) = \frac{(N-2)!}{N!}\sum_{1\le i\ne j\le N}\tilde{m}_{is}(\theta)\bar{m}_{js}(\theta)\kappa_{ij,s}$ and $\mathbb{U}_{4N}(\theta) = \frac{(N-3)!}{N!}\sum_{1\le i\ne j\ne k\le N}[\tilde{m}_{is}\bar{m}_{ks}(\theta) + \bar{m}_{is}(\theta)\tilde{m}_{ks}(\theta)]\kappa_{ij,s}$. It is easy to see that $\mathbb{U}_{3N}(\theta)$ and $\mathbb{U}_{4N}(\theta)$ are non-degenerate second and third order U-processes, respectively. One can easily apply symmetrization and similar calculations as above to verify the entropy condition in Theorem 4.10 of AC holds to conclude both $N^{1/2}\mathbb{U}_{3N}(\theta)$ and $N^{1/2}\mathbb{U}_{4N}(\theta)$ converge to Gaussian processes. This implies that

$$\begin{aligned}&\sup_{\theta\in\Theta\backslash\Theta_I}\frac{1}{N^{1/2}}\left|\tilde{S}_{Ns,3}(\theta)\right| \\ =\;&\sup_{\theta\in\Theta\backslash\Theta_I}\left|-\frac{2}{N^{3/2}}\sum_{1\le i\ne j\le N}\tilde{m}_{is}(\theta)\bar{m}_{js}(\theta)\kappa_{ij,s} + \frac{2}{N^{5/2}}\sum_{1\le i\ne j\le N}\sum_{k=1}^{N}[\tilde{m}_{is}\bar{m}_{ks}(\theta) + \bar{m}_{is}(\theta)\tilde{m}_{ks}(\theta)]\kappa_{ij,s}\right| \\ \preceq\;&\sup_{\theta\in\Theta\backslash\Theta_I}\left|N^{1/2}\mathbb{U}_{3N}(\theta)\right| + \sup_{\theta\in\Theta\backslash\Theta_I}\left|N^{1/2}\mathbb{U}_{4N}(\theta)\right| = O_p(1).\end{aligned}$$

Consequently, we have $\frac{1}{N}S_N(\theta) = \sum_{s=1}^{T-R}\sum_{\ell=1}^{3}\frac{1}{N}\tilde{S}_{Ns,\ell}(\theta) = \sum_{s=1}^{T-R}\mathrm{MDD}[m_s(Y,X,\theta)|\underline{z}_s]^2 + O_P(N^{-1/2})$ uniformly in $\theta \in \Theta\backslash\Theta_I$. ■

Proof of Theorem 3.2. In this proof Conditions (i) – (iv) listed in Lemma A.5 are referred to as C(i) – C(iv), respectively. Let $Q(\theta) \equiv \sum_{s=1}^{T-R}\mathrm{MDD}[m_s(Y,X,\theta)|\underline{z}_s]^2$ and $Q_N(\theta) \equiv \frac{1}{N}S_N(\theta) = \sum_{s=1}^{T-R}\frac{1}{N}S_{Ns}(\theta)$. Our goal is to show that, over the restricted parameter space $\Theta\cap\Theta_R$ under $\mathrm{d}_w(\cdot,\cdot)$, $Q(\cdot)$ and $Q_N(\cdot)$ as specified above satisfy C(i) - C(iv) in Lemma A.5.

We first prove that $\mathrm{d}_w(\hat{\theta}_N, \Theta_I) = O_p\left(\max\{\delta_{w,N}, N^{-1/4}\}\right) = O_p\left(N^{-1/4}\right)$ and then argue that such a rate can be improved to $O_p(\varrho_N N^{-1/2})$ by iterative arguments.

Due to the nonnegativity of MDD, $Q(\cdot) \ge 0$. By Lemma A.3, Θ is compact under $\|\cdot\|_c$ and hence is compact under $\mathrm{d}_w(\cdot,\cdot)$, which is weaker than $\|\cdot\|_c$. Since Θ_R is closed due to the continuity of $L(\cdot)$ under Assumption 2.2, $\Theta\cap\Theta_R$ is also compact under $\mathrm{d}_w(\cdot,\cdot)$. So C(i) is satisfied. Assumption 3.3(i) and (ii) guarantee C(ii) to hold with $\sigma_N = \delta_{w,N} = o\left(N^{-1/2}\right)$. C(iii) holds according to Theorem 3.1 with $b_N = N^{-1/2}$. Obviously, C(iv) holds with $a_1 = a_2 = 1$ by Lemma A.1 and the fact that $\mathrm{MDD}\left[m_s\left(Y,X,\theta^0\right)|\underline{z}_s\right] = 0$ for any $\theta^0 \in \Theta_I$. Then by Lemma A.5, we have

$$\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I\right) = O_p\left(\max\left\{\delta_{w,N}, N^{-1/4}\right\}\right) = O_p\left(N^{-1/4}\right).$$

This, in conjunction with Assumption 3.4 and Lemma A.6, implies that $\mathrm{d}_{\|\cdot\|_{L^2}}(\hat{\theta}_N, \Theta_I) = O_p(\varrho_N \, \mathrm{d}_w(\hat{\theta}_N, \Theta_I \cap \Theta_R) + \delta_{s,N}) = O_p(\varrho_N N^{-1/4})$.

For any $\epsilon > 0$, there exists a constant $K_\epsilon > 0$ such that $\Pr\left(\mathrm{d}_{\|\cdot\|_{L^2}}(\hat{\theta}_N, \Theta_I) \leq K_\epsilon \varrho_N N^{-1/4}\right) \geq 1 - \epsilon$. Let $\tilde{\Theta}_N = \{\theta : \mathrm{d}_{\|\cdot\|_{L^2}}(\theta, \Theta_I) \leq K_\epsilon \varrho_N N^{-1/4}\}$. Now, we can consider minimization of $Q_N(\theta)$ over $\theta \in \tilde{\Theta}_N$. Using arguments as used in the proof of Theorem 3.1 and the expressions in (A.31)-(A.35) in the proof of Theorem 3.3 below, we can show that

$$\sup_{\theta \in \tilde{\Theta}_N} |Q_N(\theta) - Q(\theta)| = N^{-1/2} O_p\left(\varrho_N N^{-1/4}\right) = O_p\left(\varrho_N N^{-3/4}\right) \tag{A.25}$$

by showing that $\sup_{\theta \in \tilde{\Theta}_N} \left|\frac{1}{N}\rho_{\ell N s}\left(\theta^0, \theta - \theta^0\right) - \rho_{\ell s}\left(\theta^0, \theta - \theta^0\right)\right| = O_p\left(\varrho_N N^{-3/4}\right)$ for $\ell = 1, 2, 3$ and $\sup_{\theta^0 \in \Theta_I}$ $\left|\frac{1}{N} S_N\left(\theta^0\right)\right| = O_P\left(N^{-1}\right)$. The last claim holds by Theorem 3.1(i). Next, we argue that the first claim holds for $\ell = 1$ only as the other two cases can be studied analogously. Note that with $\Delta = \theta - \theta^0$,

$$\begin{aligned}
& \frac{1}{N}\rho_{1Ns}\left(\theta^0, \theta - \theta^0\right) - \rho_{1s}\left(\theta^0, \theta - \theta^0\right) \\
= \; & -\left\{\frac{1}{N^2} \sum_{1 \leq i \neq j \leq N} \partial m_{si}[\Delta] \, \partial m_{sj}[\Delta] \, \kappa_{ij,s} - \mathbb{E}\left\{\partial m_s[\Delta] \, \partial m_s^\dagger[\Delta] \left|\underline{z}_s - \underline{z}_s^\dagger\right|\right\}\right\} \\
& + 2\left\{\frac{1}{N^3} \sum_{1 \leq i \neq j \leq N} \sum_{k=1}^N \partial m_{si}[\Delta] \, \kappa_{ij,s} \partial m_{sk}[\Delta] - \mathbb{E}\left\{\partial m_s[\Delta] \left|\underline{z}_s - \underline{z}_s^\dagger\right|\right\} \mathbb{E}\left[\partial m_s^\dagger[\Delta]\right]\right\} \\
\equiv \; & -D_{1s}(\theta) + 2D_{2s}(\theta),
\end{aligned}$$

where we suppress the dependence of $D_{1s}(\theta)$ and $D_{2s}(\theta)$ on θ^0. Let $\varkappa_s\left(\xi_i, \xi_j; \Delta\right) = \partial m_{si}[\Delta] \, \partial m_{sj}[\Delta] \, \kappa_{ij,s}$, $c_s(\Delta) = \mathbb{E}_i \mathbb{E}_j[\varkappa_s\left(\xi_i, \xi_j; \Delta\right)]$, and $\varkappa_{1s}(\xi_i; \Delta) = \mathbb{E}_j[\varkappa_s\left(\xi_i, \xi_j; \Delta\right)] - c_s(\Delta)$, where $\mathbb{E}_j$ denotes expectation w.r.t. ξ_j alone. Let $\tilde{\varkappa}_s\left(\xi_i, \xi_j; \Delta\right) = \varkappa_s\left(\xi_i, \xi_j; \Delta\right) - \varkappa_{1s}(\xi_i; \Delta) - \varkappa_{1s}\left(\xi_j; \Delta\right) + c_s(\Delta)$. By Hoeffding decomposition, we have

$$\begin{aligned}
D_{1s}(\theta) &= \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \left\{\partial m_{si}[\Delta] \, \partial m_{sj}[\Delta] \, \kappa_{ij,s} - \mathbb{E}\left[\partial m_{si}[\Delta] \, \partial m_{sj}[\Delta] \, \kappa_{ij,s}\right]\right\} + O_p\left(N^{-1}\right) \\
&= \frac{2}{N} \sum_{i=1}^N \varkappa_{1s}(\xi_i; \Delta) + \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \tilde{\varkappa}_s\left(\xi_i, \xi_j; \Delta\right) + O_p\left(N^{-1}\right),
\end{aligned} \tag{A.26}$$

where $O_p\left(N^{-1}\right)$ holds uniformly in θ (and θ^0). Noting that the second term in (A.26) is a degenerate second order U-process, we can readily follow the proof of Theorem 3.1(i) and show that it is $O_P\left(N^{-1}\right)$ uniformly in $\Delta = \theta - \theta^0 \in \Theta - \theta^0$. For the first term in (A.26), we can apply the expression of $\frac{\partial m_s\left(Y, X, \theta^0\right)}{\partial \theta}[\Delta]$ in (A.35) and entropy calculations as used in the proof of Theorem 3.1(i) to show that

$$\begin{aligned}
& \sup_{\theta \in \tilde{\Theta}_N} \left|\frac{1}{N} \sum_{i=1}^N \varkappa_{1s}\left(\xi_i; \theta - \theta^0\right)\right| \\
= \; & \sup_{\theta \in \tilde{\Theta}_N} \left|\frac{1}{N} \sum_{i=1}^N \left\{\mathbb{E}_j[\partial m_{si}\left[\theta - \theta^0\right] \partial m_{sj}\left[\theta - \theta^0\right] \kappa_{ij,s}] - \mathbb{E}_i \mathbb{E}_j[\partial m_{si}\left[\theta - \theta^0\right] \partial m_{sj}\left[\theta - \theta^0\right] \kappa_{ij,s}]\right\}\right| \\
= \; & N^{-1/2} O_p\left(\varrho_N N^{-1/4}\right) = O_p\left(\varrho_N N^{-3/4}\right).
\end{aligned}$$

Then $\sup_{\theta \in \tilde{\Theta}_N} |D_{1s}(\theta)| = O_p\left(\varrho_N N^{-3/4}\right)$. Analogously, we can show that $\sup_{\theta \in \tilde{\Theta}_N} |D_{2s}(\theta)| = O_p\left(\varrho_N N^{-3/4}\right)$. It follows that $\sup_{\theta \in \tilde{\Theta}_N} \left|\frac{1}{N}\rho_{1Ns}\left(\theta^0, \theta - \theta^0\right) - \rho_{1s}\left(\theta^0, \theta - \theta^0\right)\right| = O_p\left(\varrho_N N^{-3/4}\right)$ and (A.25) holds. Then we can apply Lemma A.5 with $b_N = \varrho_N N^{-3/4}$ to conclude

$$\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I\right) = O_p\left(\max\left\{\delta_{w,N}, \varrho_N^{1/2} N^{-3/8}\right\}\right) = O_p(\varrho_N^{1/2} N^{-3/8}). \tag{A.27}$$

Now, given the first iteration result in (A.27), we can focus on $\tilde{\Theta}_N^{(1)} = \{\theta : \mathrm{d}_{\|\cdot\|_{L^2}}(\theta, \Theta_I) \leq K_\epsilon \varrho_N \varrho_N^{1/2} N^{-3/8}\}$ and show that

$$\sup_{\theta\in\tilde{\Theta}_N^{(1)}} |Q_N(\theta) - Q(\theta)| = N^{-1/2} O_p\left(\varrho_N \varrho_N^{1/2} N^{-3/8}\right) = O_p\left(\varrho_N^{3/2} N^{-7/8}\right)$$

which, in conjunction with Lemma A.5 implies that

$$\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I\right) = O_p\left(\max\left\{\delta_{w,N}, \varrho_N^{3/4} N^{-7/16}\right\}\right) = O_p(\varrho_N^{3/4} N^{-7/16}). \tag{A.28}$$

Repeating such an arguments for any finite m times, we can obtain

$$\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I\right) = N^{-1/4} O_p\left(\left(N^{-1/4}\varrho_N\right)^{\sum_{j=1}^m \frac{1}{2^j}}\right).$$

By choosing $m = m(\epsilon)$ sufficiently large, we obtain $\mathrm{d}_w(\hat{\theta}_N, \Theta_I) = o_p(N^{-\frac{1}{2}+\frac{\epsilon}{4}} \varrho_N^{1-\epsilon})$ for any fixed small positive number $\epsilon > 0$. ■

Proof of Theorem 3.3. We organize this proof into three parts. In Part I, we establish that, to calculate the test statistic, we only need to minimize over a neighborhood of $\Theta_I \cap \Theta_R$; in Part II, we show further that the minimization would focus on $\Theta_I \cap \Theta_R$ asymptotically; and in Part III, we combine the results in Parts I and II to complete the proof. Note that continuity and compactness imply that all minimums are indeed attained.

Part I. Recall that $\hat{\theta}_N \in \underset{\theta\in\Theta_N\cap\Theta_R}{\operatorname{argmin}} S_N(\theta)$. Then $\mathrm{d}_{\|\cdot\|_{L^2}}\left(\hat{\theta}_N, \Theta_I\cap\Theta_R\right) = O_p\left(\varrho_N \,\mathrm{d}_w(\hat{\theta}_N, \Theta_I) + \delta_{s,N}\right) = O_p(\varrho_N o(\varrho_N^{1-\epsilon} N^{-\frac{1}{2}+\frac{\epsilon}{4}}) + \delta_{s,N}) = o_p(N^{-1/4})$ by Lemma A.6, Theorem 3.2, and Assumption 3.4. This, in conjunction with the fact that $\delta_{s,N} = \sup_{\theta\in\Theta_I\cap\Theta_R} \|\Pi_N\theta - \theta\|_{L^2} = o\left(N^{-1/4}\right)$ by Assumption 3.3(i), implies that $\exists$ some $\delta_N \downarrow 0$

$$\delta_N = o\left(N^{-1/4}\right), \;\; \mathrm{d}_{\|\cdot\|_{L^2}}\left(\hat{\theta}_N, \Theta_I\cap\Theta_R\right) = o_p(\delta_N), \;\; \max\left\{\delta_{s,N}, N^{-1/2}\right\} = o(\delta_N). \tag{A.29}$$

Define $B_N^{\delta_N}(\theta^0) \equiv \left\{\theta \in \Theta_N \cap \Theta_R : \left\|\theta - \theta^0\right\|_{L^2} \leq \delta_N\right\}$. Then (A.29) implies that

$$\min_{\theta\in\Theta_N\cap\Theta_R} S_N(\theta) = \inf_{\theta^0\in\Theta_I\cap\Theta_R}\left[\min_{\theta\in B_N^{\delta_N}(\theta^0)} S_N(\theta)\right] + o_p(1) \tag{A.30}$$

because $\forall\, \varepsilon > 0$,

$$\begin{aligned}
& \Pr\left\{\left|\min_{\theta\in\Theta_N\cap\Theta_R} S_N(\theta) - \inf_{\theta^0\in\Theta_I\cap\Theta_R}\left[\min_{\theta\in B_N^{\delta_N}(\theta^0)} S_N(\theta)\right]\right| > \varepsilon\right\} \\
= \;& \Pr\left\{\left|S_N\left(\hat{\theta}_N\right) - \inf_{\theta^0\in\Theta_I\cap\Theta_R}\left[\min_{\theta\in B_N^{\delta_N}(\theta^0)} S_N(\theta)\right]\right| > \varepsilon\right\} \\
\leq \;& \Pr\left\{\hat{\theta}_N \notin \bigcup_{\theta^0\in\Theta_I\cap\Theta_R} B_N^{\delta_N}(\theta^0)\right\} \\
= \;& \Pr\left\{\mathrm{d}_{\|\cdot\|_{L^2}}\left(\hat{\theta}_N, \Theta_I\cap\Theta_R\right) > \delta_N\right\} \to 0,
\end{aligned}$$

where the convergence follows from the second condition in (A.29).

Continuing with (A.30),

$$
\begin{aligned}
& \min_{\theta\in\Theta_N\cap\Theta_R} S_N(\theta) \\
= & \inf_{\theta^0\in\Theta_I\cap\Theta_R}\left[\min_{\theta\in B_N^{\delta_N}(\theta^0)}\sum_{s=1}^{T-R} S_{Ns}(\theta)\right]+o_p(1) \\
= & \inf_{\theta^0\in\Theta_I\cap\Theta_R}\left[\min_{\theta\in B_N^{\delta_N}(\theta^0)}\sum_{s=1}^{T-R}\left[S_{Ns}\left(\theta^0\right)+\rho_{1Ns}\left(\theta^0,\Delta\right)+\rho_{2Ns}\left(\theta^0,\Delta\right)+\rho_{3Ns}\left(\theta^0,\Delta\right)\right]\right]+o_p(1), \qquad \text{(A.31)}
\end{aligned}
$$

where $\Delta=\theta-\theta^0$,

$$
\rho_{1Ns}\left(\theta^0,\Delta\right) \equiv -\frac{1}{N}\sum_{1\le i\neq j\le N}\partial m_{si}[\Delta]\,\partial m_{sj}[\Delta]\,\kappa_{ij,s}+\frac{2}{N^2}\sum_{1\le i\neq j\le N}\sum_{k=1}^{N}\partial m_{si}[\Delta]\,\kappa_{ij,s}\partial m_{sk}[\Delta], \qquad \text{(A.32)}
$$

$$
\begin{aligned}
\rho_{2Ns}\left(\theta^0,\Delta\right) \equiv & -\frac{1}{N}\sum_{1\le i\neq j\le N}\partial m_{si}[\Delta]\,\partial^2 m_{sj}[\Delta]\,\kappa_{ij,s} \\
& +\frac{1}{N^2}\sum_{1\le i\neq j\le N}\sum_{k=1}^{N}\left\{\partial m_{si}[\Delta]\,\partial^2 m_{sk}[\Delta]+\partial^2 m_{si}[\Delta]\,\partial m_{sk}[\Delta]\right\}\kappa_{ij,s}, \qquad \text{(A.33)}
\end{aligned}
$$

$$
\rho_{3Ns}\left(\theta^0,\Delta\right) \equiv \frac{-1}{4N}\sum_{1\le i\neq j\le N}\partial^2 m_{si}[\Delta]\,\partial^2 m_{sj}[\Delta]\,\kappa_{ij,s}+\frac{1}{2N^2}\sum_{1\le i\neq j\le N}\sum_{k=1}^{N}\partial^2 m_{si}[\Delta]\,\kappa_{ij,s}\partial^2 m_{sk}[\Delta], \qquad \text{(A.34)}
$$

with $\partial m_{si}[\Delta]=\frac{\partial m_s\left(Y_i,X_i,\theta^0\right)}{\partial\theta}[\Delta]$ and $\partial^2 m_{si}[\Delta]=\frac{\partial^2 m_s\left(Y_i,X_i,\theta^0\right)}{\partial\theta^2}[\Delta,\Delta]$. Note that the last equality in (A.31) above is obtained by plugging

$$
m_s(Y,X,\theta)=m_s\left(Y,X,\theta^0\right)+\frac{\partial m_s\left(Y,X,\theta^0\right)}{\partial\theta}[\Delta]+\frac{1}{2}\frac{\partial^2 m_s\left(Y,X,\theta^0\right)}{\partial\theta^2}[\Delta,\Delta]
$$

(with $\Delta=\theta-\theta^0$) into $S_{Ns}(\theta)=-\frac{1}{N}\sum_{1\le i\neq j\le N}m_s(Y_i,X_i,\theta)\,m_s(Y_j,X_j,\theta)\,\kappa_{ij,s}+\frac{2}{N^2}\sum_{1\le i\neq j\le N}m_s(Y_i,X_i,\theta)$ $\times\kappa_{ij,s}\sum_{k=1}^{N}m_s(Y_k,X_k,\theta)$.

Let $\partial m_s[\Delta]=\frac{\partial m_s\left(Y,X,\theta^0\right)}{\partial\theta}[\Delta]$, $\partial m_s^\dagger[\Delta]=\frac{\partial m_s\left(Y^\dagger,X^\dagger,\theta^0\right)}{\partial\theta}[\Delta]$, $\partial^2 m_s[\Delta]=\frac{\partial^2 m_s\left(Y,X,\theta^0\right)}{\partial\theta^2}[\Delta,\Delta]$ and $\partial^2 m_s^\dagger[\Delta]$ $=\frac{\partial^2 m_s\left(Y^\dagger,X^\dagger,\theta^0\right)}{\partial\theta^2}[\Delta,\Delta]$. The population analogue of $N^{-1}\rho_{1Ns}\left(\theta^0,\Delta\right)$, $N^{-1}\rho_{2Ns}\left(\theta^0,\Delta\right)$ and $N^{-1}\rho_{3Ns}\left(\theta^0,\Delta\right)$ are respectively given by

$$
\begin{aligned}
\rho_{1s}\left(\theta^0,\Delta\right) \equiv & -\mathbb{E}\left\{\partial m_s[\Delta]\,\partial m_s^\dagger[\Delta]\left|\underline{z}_s-\underline{z}_s^\dagger\right|\right\}+2\mathbb{E}\left\{\partial m_s[\Delta]\left|\underline{z}_s-\underline{z}_s^\dagger\right|\right\}\mathbb{E}\left[\partial m_s^\dagger[\Delta]\right], \\
\rho_{2s}\left(\theta^0,\Delta\right) \equiv & -\mathbb{E}\left\{\partial m_s[\Delta]\,\partial^2 m_s^\dagger[\Delta]\left|\underline{z}_s-\underline{z}_s^\dagger\right|\right\}+\mathbb{E}\left\{\partial m_s[\Delta]\left|\underline{z}_s-\underline{z}_s^\dagger\right|\right\}\mathbb{E}\left[\partial^2 m_s^\dagger[\Delta]\right] \\
& +\mathbb{E}\left\{\partial^2 m_s[\Delta]\left|\underline{z}_s-\underline{z}_s^\dagger\right|\right\}\mathbb{E}\left[\partial m_s^\dagger[\Delta]\right], \text{ and} \\
\rho_{3s}\left(\theta^0,\Delta\right) \equiv & -\frac{1}{4}\mathbb{E}\left\{\partial^2 m_s[\Delta]\,\partial^2 m_s^\dagger[\Delta]\left|\underline{z}_s-\underline{z}_s^\dagger\right|\right\}+\frac{1}{2}\mathbb{E}\left\{\partial^2 m_s[\Delta]\left|\underline{z}_s-\underline{z}_s^\dagger\right|\right\}\mathbb{E}\left[\partial^2 m_s^\dagger[\Delta]\right].
\end{aligned}
$$

By straightforward pathwise derivative calculations,

$$
\begin{aligned}
\frac{\partial m_s\left(Y,X,\theta^0\right)}{\partial\theta}[\Delta] = & \sum_{t=1}^{R}\left(\phi_{s,t}-\phi_{s,t}^0\right)\left[y_{T-R+t}-g^0\left(x_{T-R+t}\right)\right]-\left[g\left(x_s\right)-g^0\left(x_s\right)\right] \\
& -\sum_{t=1}^{R}\phi_{s,t}^0\left[g\left(x_{T-R+t}\right)-g^0\left(x_{T-R+t}\right)\right], \\
\frac{\partial^2 m_s\left(Y,X,\theta^0\right)}{\partial\theta^2}[\Delta,\Delta] = & -2\sum_{t=1}^{R}\left(\phi_{s,t}-\phi_{s,t}^0\right)\left[g\left(x_{T-R+t}\right)-g^0\left(x_{T-R+t}\right)\right]. \qquad \text{(A.35)}
\end{aligned}
$$

It is easy to verify that $\sum_{\ell=1}^{3} \rho_{\ell s}\left(\theta^{0}, \theta-\theta^{0}\right)=\operatorname{MDD}\left[m_{s}(Y, X, \theta)-m_{s}\left(Y, X, \theta^{0}\right) \mid \underline{z}_{s}\right]^{2}=\operatorname{MDD}\left[m_{s}(Y, X, \theta) \mid \underline{z}_{s}\right]^{2}$, where the last equality follows from Lemma A.1.

Using arguments analogous to those in the proof of Theorem 3.1(ii), we can show that for any given $\theta^{0} \in \Theta_{I}$,

$$\frac{1}{N} \rho_{1 N s}\left(\theta^{0}, \Delta\right)=\rho_{1 s}\left(\theta^{0}, \Delta\right)+O_{p}\left(N^{-1 / 2}\right)$$

uniformly in $\Delta=\theta-\theta^{0} \in \Theta-\theta^{0} \equiv\left\{\tilde{\theta}-\theta^{0}: \tilde{\theta} \in \Theta\right\}$. It follows that $\sqrt{N}\left[\frac{1}{N} \rho_{1 N s}\left(\theta^{0}, \Delta\right)-\rho_{1 s}\left(\theta^{0}, \Delta\right)\right]=O_{p}(1)$ uniformly in $\Delta=\theta-\theta^{0} \in \Theta-\theta^{0}$, which in turn implies the uniform asymptotic $\|\cdot\|_{L^{2}}$-equicontinuity in probability in the sense that $\forall\ \epsilon_{N} \downarrow 0$, it holds that

$$\sup_{\left\|\theta_{1}-\theta_{2}\right\|_{L^{2}} \leq \epsilon_{N}} \sqrt{N}\left|\left[\frac{1}{N} \rho_{1 N s}\left(\theta^{0}, \theta_{1}\right)-\rho_{1 s}\left(\theta^{0}, \theta_{1}\right)\right]-\left[\frac{1}{N} \rho_{1 N s}\left(\theta^{0}, \theta_{2}\right)-\rho_{1 s}\left(\theta^{0}, \theta_{2}\right)\right]\right|=o_{p}(1). \tag{A.36}$$

It is easy to see algebraically that $a \rho_{1 N s}\left(\theta^{0}, \Delta\right)=\rho_{1 N s}\left(\theta^{0}, \sqrt{a} \Delta\right)$ and $a \rho_{1 s}\left(\theta^{0}, \Delta\right)=\rho_{1 s}\left(\theta^{0}, \sqrt{a} \Delta\right)$ for any scalar constant $a \geq 0$. These algebraic properties imply that

$$\begin{aligned}
& \sup_{\left\|\theta_{1}-\theta_{2}\right\|_{L^{2}} \leq \delta_{N}}\left|\left[\rho_{1 N s}\left(\theta^{0}, \theta_{1}\right)-N \rho_{1 s}\left(\theta^{0}, \theta_{1}\right)\right]-\left[\rho_{1 N s}\left(\theta^{0}, \theta_{2}\right)-N \rho_{1 s}\left(\theta^{0}, \theta_{2}\right)\right]\right| \\
= & \sup_{\left\|N^{\frac{1}{4}} \theta_{1}-N^{\frac{1}{4}} \theta_{2}\right\|_{L^{2}} \leq N^{\frac{1}{4}} \delta_{N}} \sqrt{N}\left|\left[\frac{1}{N} \rho_{1 N s}(\theta^{0}, N^{\frac{1}{4}} \theta_{1})-\rho_{1 s}(\theta^{0}, N^{\frac{1}{4}} \theta_{1})\right]-\left[\frac{1}{N} \rho_{1 N s}(\theta^{0}, N^{\frac{1}{4}} \theta_{2})-\rho_{1 s}(\theta^{0}, N^{\frac{1}{4}} \theta_{2})\right]\right| \\
= & o_{p}(1),
\end{aligned} \tag{A.37}$$

where the last equality holds because $N^{\frac{1}{4}} \delta_{N} \downarrow 0$ according to (A.29). This, in conjunction with the fact that $\rho_{1 N s}\left(\theta^{0}, 0\right)=\rho_{1 s}\left(\theta^{0}, 0\right)=0$, implies that

$$\sup_{\theta:\left\|\theta-\theta^{0}\right\|_{L^{2}} \leq \delta_{N}}\left|\rho_{1 N s}\left(\theta^{0}, \theta-\theta^{0}\right)-N \rho_{1 s}\left(\theta^{0}, \theta-\theta^{0}\right)\right|=o_{p}(1). \tag{A.38}$$

Analogously to the analysis of $\rho_{1 N s}\left(\theta^{0}, \Delta\right)$, it can be shown that

$$\begin{aligned}
\sqrt{N}\left[\frac{1}{N} \rho_{2 N s}\left(\theta^{0}, \Delta\right)-\rho_{2 s}\left(\theta^{0}, \Delta\right)\right] & = O_{p}(1) \text{ uniformly in } \Delta=\theta-\theta^{0} \in \Theta-\theta^{0} \text{ and} \\
\sup_{\theta:\left\|\theta-\theta^{0}\right\|_{L^{2}} \leq \delta_{N}}\left|\rho_{2 N s}\left(\theta^{0}, \theta-\theta^{0}\right)-N \rho_{2 s}\left(\theta^{0}, \theta-\theta^{0}\right)\right| & = o_{p}(1),
\end{aligned} \tag{A.39}$$

where we use the fact that $a \rho_{2 N s}\left(\theta^{0}, \Delta\right)=\rho_{2 N s}\left(\theta^{0}, a^{1 / 3} \Delta\right)$ and $a \rho_{2 s}\left(\theta^{0}, \Delta\right)=\rho_{2 s}\left(\theta^{0}, a^{1 / 3} \Delta\right)$ for any scalar constant $a \geq 0$, $\rho_{2 N s}\left(\theta^{0}, 0\right)=\rho_{2 s}\left(\theta^{0}, 0\right)=0$, and that $N^{\frac{1}{6}} \delta_{N} \downarrow 0$. Similarly, it can be shown that

$$\begin{aligned}
\sqrt{N}\left[\frac{1}{N} \rho_{3 N s}\left(\theta^{0}, \Delta\right)-\rho_{3 s}\left(\theta^{0}, \Delta\right)\right] & = O_{p}(1) \text{ uniformly in } \Delta=\theta-\theta^{0} \in \Theta-\theta^{0} \text{ and} \\
\sup_{\theta:\left\|\theta-\theta^{0}\right\|_{L^{2}} \leq \delta_{N}}\left|\rho_{3 N s}\left(\theta^{0}, \theta-\theta^{0}\right)-N \rho_{3 s}\left(\theta^{0}, \theta-\theta^{0}\right)\right| & = o_{p}(1),
\end{aligned} \tag{A.40}$$

where we use the fact that that $a \rho_{3 N s}\left(\theta^{0}, \Delta\right)=\rho_{3 N s}\left(\theta^{0}, a^{1 / 4} \Delta\right)$ and $a \rho_{3 s}\left(\theta^{0}, \Delta\right)=\rho_{3 s}\left(\theta^{0}, a^{1 / 4} \Delta\right)$ for any scalar constant $a \geq 0$, $\rho_{3 N s}\left(\theta^{0}, 0\right)=\rho_{3 s}\left(\theta^{0}, 0\right)=0$, and that $N^{\frac{1}{8}} \delta_{N} \downarrow 0$. Combining (A.38), (A.39), and (A.40), we have

$$\sum_{\ell=1}^{3} \rho_{\ell N s}\left(\theta^{0}, \theta-\theta^{0}\right)=N \sum_{\ell=1}^{3} \rho_{\ell s}\left(\theta^{0}, \theta-\theta^{0}\right)+o_{p}(1)=N \cdot \operatorname{MDD}\left[m_{s}(Y, X, \theta) \mid \underline{z}_{s}\right]^{2}+o_{p}(1), \tag{A.41}$$

where each $o_p(1)$ holds uniformly in $\theta \in B_N^{\delta_N}(\theta^0)$.

Plugging (A.41) in (A.31) yields

$$\begin{aligned}
\min_{\theta\in\Theta_N\cap\Theta_R} S_N(\theta) &= \inf_{\theta^0\in\Theta_I\cap\Theta_R}\left[\min_{\theta\in B_N^{\delta_N}(\theta^0)}\sum_{s=1}^{T-R}\left[S_{Ns}\left(\theta^0\right)+N\cdot\mathrm{MDD}\left[m_s(Y,X,\theta)\,|\underline{z}_s\right]^2\right]\right]+o_p(1)\\
&= \inf_{\theta^0\in\Theta_I\cap\Theta_R}\left[S_N\left(\theta^0\right)+\min_{\theta\in B_N^{\delta_N}(\theta^0)} N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s(Y,X,\theta)\,|\underline{z}_s\right]^2\right]+o_p(1). \quad \text{(A.42)}
\end{aligned}$$

Part II. Next, we show that the term $\min_{\theta\in B_N^{\delta_N}(\theta^0)} N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s(Y,X,\theta)\,|\underline{z}_s\right]^2$ in (A.42) can be dropped asymptotically. The nonnegativity of MDD implies that

$$\min_{\theta\in B_N^{\delta_N}(\theta^0)} N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s(Y,X,\theta)\,|\underline{z}_s\right]^2 \geq 0. \quad \text{(A.43)}$$

According to the third condition in (A.29), i.e., $\sup_{\theta\in\Theta_I\cap\Theta_R}\|\Pi_N\theta-\theta\|_{L^2}=o(\delta_N)$, it holds that $\left\|\Pi_N\theta^0-\theta^0\right\|_{L^2}\leq\delta_N$, or equivalently, $\Pi_N\theta^0\in B_N^{\delta_N}\left(\theta^0\right)$for all $\theta^0\in\Theta_I\cap\Theta_R$ and large enough N. Therefore,

$$\min_{\theta\in B_N^{\delta_N}(\theta^0)} N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s(Y,X,\theta)\,|\underline{z}_s\right]^2 \;\leq\; N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\Pi_N\theta^0\right)|\underline{z}_s\right]^2 \quad \text{(A.44)}$$

for all $\theta^0\in\Theta_I\cap\Theta_R$ and large enough N. Then for all $\theta^0\in\Theta_I\cap\Theta_R$,

$$\begin{aligned}
0 &\leq N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\Pi_N\theta^0\right)|\underline{z}_s\right]^2\\
&= N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\Pi_N\theta^0\right)-m_s\left(Y,X,\theta^0\right)|\underline{z}_s\right]^2\\
&= N\,\mathrm{d}_w\left(\Pi_N\theta^0,\theta^0\right)^2 \leq \left[N^{1/2}\sup_{\theta^0\in\Theta_I\cap\Theta_R}\mathrm{d}_w\left(\Pi_N\theta^0,\theta^0\right)\right]^2 = o(1),
\end{aligned}$$

where the first equality follows from Lemma A.1 and the fact that $\mathrm{MDD}\left[m_s\left(Y,X,\theta^0\right)|\underline{z}_s\right]^2=0$, and the last equality follows from Assumption 3.3(ii). It follows that

$$\sup_{\theta^0\in\Theta_I\cap\Theta_R} N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\Pi_N\theta^0\right)|\underline{z}_s\right]^2 \;=\; o(1). \quad \text{(A.45)}$$

Combining (A.43)–(A.45) yields

$$\sup_{\theta^0\in\Theta_I\cap\Theta_R}\min_{\theta\in B_N^{\delta_N}(\theta^0)} N\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s(Y,X,\theta)\,|\underline{z}_s\right]^2 \;=\; o(1) \quad \text{(A.46)}$$

As a result, substituting (A.46) into (A.42) delivers

$$\begin{aligned}
\min_{\theta\in\Theta_N\cap\Theta_R} S_N(\theta) &= \inf_{\theta^0\in\Theta_I\cap\Theta_R}\left\{S_N\left(\theta^0\right)+\min_{\theta\in B_N^{\delta_N}(\theta^0)} N\left[\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s(Y,X,\theta)\,|\underline{z}_s\right]^2\right]\right\}+o_p(1)\\
&= \inf_{\theta^0\in\Theta_I\cap\Theta_R} S_N\left(\theta^0\right)+o_p(1). \quad \text{(A.47)}
\end{aligned}$$

Part III. By Theorem 3.1(i) and the extended continuous mapping (see, e.g., Theorem 1.11.1 in van der Vaart and Wellner (1996)), we have

$$\inf_{\theta^0\in\Theta_I\cap\Theta_R}\left[\sum_{s=1}^{T-R} S_{Ns}\left(\theta^0\right)\right]\xrightarrow{\mathcal{L}}\inf_{\theta^0\in\Theta_I\cap\Theta_R}\sum_{s=1}^{T-R}\left[\mathbb{B}_s\left(\theta^0\right)+\mathbb{C}_s\left(\theta^0\right)\right],$$

which, in conjunction with (A.47), implies that $\min_{\theta\in\Theta_N\cap\Theta_R} S_N\left(\theta\right)\xrightarrow{\mathcal{L}}\inf_{\theta^0\in\Theta_I\cap\Theta_R}\sum_{s=1}^{T-R}\left[\mathbb{B}_s\left(\theta^0\right)+\mathbb{C}_s\left(\theta^0\right)\right].$ So the proof is completed. ■

Proof of Theorem 3.4. Let δ_N be the same as the one specified by (A.29) in the proof of Theorem 3.3. By Theorem 3.1,

$$\sup_{\theta\in\Theta}\left|\frac{1}{N}S_N\left(\theta\right)-\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\theta\right)|\underline{z}_s\right]^2\right|\xrightarrow{p}0. \tag{A.48}$$

Since Θ is compact under $\left\|\cdot\right\|_c$ by Lemma A.3, it follows from the theorem of maximum and the continuous mapping theorem that

$$\min_{\theta\in\Theta\cap\Theta_R}\frac{1}{N}S_N\left(\theta\right)=\min_{\theta\in\Theta\cap\Theta_R}\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\theta\right)|\underline{z}_s\right]^2+o_p\left(1\right). \tag{A.49}$$

Let $\tilde{\theta}_N\in\operatorname*{argmin}_{\theta\in\Theta\cap\Theta_R} S_N\left(\theta\right)$. Note that

$$\begin{aligned}
0 &\le \min_{\theta\in\Theta_N\cap\Theta_R}\frac{1}{N}S_N\left(\theta\right)-\min_{\theta\in\Theta\cap\Theta_R}\frac{1}{N}S_N\left(\theta\right)\\
&\le \frac{1}{N}S_N\left(\Pi_N\tilde{\theta}_N\right)-\frac{1}{N}S_N\left(\tilde{\theta}_N\right)\\
&= \left|\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\Pi_N\tilde{\theta}_N\right)|\underline{z}_s\right]^2-\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\tilde{\theta}_N\right)|\underline{z}_s\right]^2\right|+o_p\left(1\right)\\
&\preceq \sum_{s=1}^{T-R}\left\{\mathrm{MDD}\left[m_s\left(Y,X,\Pi_N\tilde{\theta}_N\right)-m_s\left(Y,X,\tilde{\theta}_N\right)|\underline{z}_s\right]^2\right\}^{1/2}+o_p\left(1\right)\\
&\le \left(T-R\right)^{1/2}\left\{\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\Pi_N\tilde{\theta}_N\right)-m_s\left(Y,X,\tilde{\theta}_N\right)|\underline{z}_s\right]^2\right\}^{1/2}+o_p\left(1\right)\\
&\preceq \left\|\Pi_N\tilde{\theta}_N-\tilde{\theta}_N\right\|_{L^2}+o_p\left(1\right)\\
&= o_p\left(1\right),
\end{aligned} \tag{A.50}$$

where the first equality holds because of the uniform asymptotic $\left\|\cdot\right\|_{L^2}$-equicontinuity of $\frac{1}{N}S_N\left(\theta\right)-\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\theta\right)|\underline{z}_s\right]^2$ (implied by (A.48) or by Theorem 3.1 directly) and the fact that $\left\|\Pi_N\tilde{\theta}_N-\tilde{\theta}_N\right\|_{L^2}=o\left(\delta_N\right)$ with $\delta_N\downarrow 0$, the third inequality follows from Lemma A.2,[12] the fourth inequality holds by Cauchy-Schwarz inequality, and the last inequality holds by Lemma 3.1.

(A.50) and (A.49) imply

$$\frac{1}{N}\hat{S}_N=\min_{\theta\in\Theta_N\cap\Theta_R}\frac{1}{N}S_N\left(\theta\right)=\min_{\theta\in\Theta\cap\Theta_R}\sum_{s=1}^{T-R}\mathrm{MDD}\left[m_s\left(Y,X,\theta\right)|\underline{z}_s\right]^2+o_p\left(1\right).$$

[12] The finite moments requirement in Lemma A.2 can be easily verified by Assumption 3.1(ii), together with the result that $\left|m_s\left(Y,X,\theta\right)\right|\preceq\left|Y\right|+1$, which is established in (A.20) in the proof of Theorem 3.1.

If $\Theta_I \cap \Theta_R = \emptyset$, then $\forall\ \theta \in \Theta \cap \Theta_R$ it holds that $\sum_{s=1}^{T-R} \mathrm{MDD}\left[m_s\left(Y,X,\theta\right)|\underline{z}_s\right]^2 > 0$. The compactness of $\Theta \cap \Theta_R$ guarantees that

$$\min_{\theta \in \Theta \cap \Theta_R} \sum_{s=1}^{T-R} \mathrm{MDD}\left[m_s\left(Y,X,\theta\right)|\underline{z}_s\right]^2 > 0,$$

which completes the proof.[13] ■

Proof of Theorem 3.5. Recall that $m_{is}^*\left(\theta\right) = m_{is}\left(\theta\right) v_i$ Note that $S_{Ns}^*\left(\theta\right) = S_{Ns,1}^*\left(\theta\right) + S_{Ns,2}^*\left(\theta\right)$, where $S_{Ns,1}^*\left(\theta\right) = -\frac{1}{N}\sum_{1\le i\ne j\le N} m_{is}^*\left(\theta\right) m_{js}^*\left(\theta\right)\kappa_{ij,s}$ and $S_{Ns,2}^*\left(\theta\right) = \frac{2}{N}\sum_{1\le i\ne j\le N} m_{is}^*\left(\theta\right)\kappa_{ij,s}\frac{1}{N}\sum_{k=1}^{N} m_{ks}^*\left(\theta\right)$. Note that

$$S_{Ns,2}^*\left(\theta\right) = \frac{2}{N^2}\sum_{1\le i\ne j\le N} m_{is}^*\left(\theta\right)^2\kappa_{ij,s} + \frac{2}{N^2}\sum_{1\le i\ne j\le N} m_{is}^*\left(\theta\right) m_{js}^*\left(\theta\right)\kappa_{ij,s} + \frac{(N-1)(N-2)}{N^2} N\mathbb{U}_{2Ns}^*,$$

where $\mathbb{U}_{2Ns}^* = \binom{N}{3}^{-1}\sum_{1\le i<j\le k\le N}\psi_s\left(\xi_i^*,\xi_j^*,\xi_k^*;\theta\right)$, $\xi_i^* = (\xi_i', v_i)'$ and $\psi_s^*(\xi_i^*,\xi_j^*,\xi_k^*;\theta) = \frac{1}{3}[m_{is}^*\left(\theta\right) m_{ks}^*\left(\theta\right)\kappa_{ij,s}$

$+m_{is}^*\left(\theta\right) m_{js}^*\left(\theta\right)\kappa_{ik,s} + m_{js}^*\left(\theta\right) m_{ks}^*\left(\theta\right)\kappa_{jk,s} + m_{js}^*\left(\theta\right) m_{is}^*\left(\theta\right)\kappa_{jk,s} + m_{ks}^*\left(\theta\right) m_{is}^*\left(\theta\right)\kappa_{jk,s} + m_{ks}^*\left(\theta\right) m_{js}^*\left(\theta\right)\kappa_{ik,s}]$ is a symmetrized version of $\psi_{0s}\left(\xi_i,\xi_j,\xi_k;\theta\right) \equiv 2m_{is}^*\left(\theta\right) m_{ks}^*\left(\theta\right)\kappa_{ij,s}$. Note that

$$\begin{aligned}
\mathbb{E}^*\left[\psi_s\left(\xi_1^*,\xi_2^*,\xi_3^*;\theta\right)\right] &= 0,\ \mathbb{E}^*\left[\psi_s\left(\xi_1^*,\xi_2^*,\xi_3^*;\theta\right)|\xi_1^*\right] = 0 \text{ and} \\
\mathbb{E}^*\left[\psi_s\left(\xi_1^*,\xi_2^*,\xi_3^*;\theta\right)|\xi_1^*,\xi_2^*\right] &= \frac{1}{3} m_{1s}^*\left(\theta\right) m_{2s}^*\left(\theta\right)\mathbb{E}_3(\kappa_{13,s}+\kappa_{23,s}) \equiv h_s^{(2)}\left(\xi_1^*,\xi_2^*;\theta\right).
\end{aligned}$$

Let $h_s^{(3*)}\left(\xi_1^*,\xi_2^*,\xi_3^*;\theta\right) = \psi_s\left(\xi_1^*,\xi_2^*,\xi_3^*;\theta\right) - [h_s^{(2)}\left(\xi_1^*,\xi_2^*;\theta\right) + h_s^{(2)}\left(\xi_1^*,\xi_3^*;\theta\right) + h_s^{(2)}\left(\xi_2^*,\xi_3^*;\theta\right)]$. By Hoeffding's decomposition in (A.13), we have $\mathbb{U}_{2Ns}^*\left(\theta\right) = 3\mathbb{H}_{2Ns}^*\left(\theta\right) + \mathbb{H}_{3Ns}^*\left(\theta\right)$, where

$$\mathbb{H}_{2Ns}\left(\theta\right) = \binom{N}{3}^{-1}\sum_{1\le i<j\le N} h_s^{(2*)}\left(\xi_i^*,\xi_j^*;\theta\right) \text{ and } \mathbb{H}_{3Ns}\left(\theta\right) = \binom{N}{3}^{-1}\sum_{1\le i<j\le k\le N} h_s^{(3*)}\left(\xi_i^*,\xi_j^*,\xi_k^*;\theta\right),$$

where $h_s^{(2*)}\left(\xi_i^*,\xi_j^*;\theta\right) = \frac{1}{N}\sum_{k=1}^{N} h_s^{(2)}(\xi_i^*,\xi_j^*;\underline{z}_{k,s},\theta) = \frac{1}{3} m_{is}^*\left(\theta\right) m_{js}^*\left(\theta\right)\mathbb{E}_k(\kappa_{ik,s}+\kappa_{jk,s})$. Similarly, $S_{Ns,1}^*\left(\theta\right) = \frac{N-1}{N} N\mathbb{U}_{1Ns}^*\left(\theta\right)$, where $\mathbb{U}_{1N}^*\left(\theta\right) = -\binom{N}{2}^{-1}\sum_{1\le i\ne j\le N} m_{is}^*\left(\theta\right) m_{js}^*\left(\theta\right)\kappa_{ij,s}$. Then we have

$$S_{Ns}^*\left(\theta\right) = N\mathbb{U}_{Ns}^*\left(\theta\right) + N\mathbb{H}_{3Ns}^*\left(\theta\right) + \frac{2}{N^2}\sum_{1\le i\ne j\le N} m_{is}^*\left(\theta\right)^2\kappa_{ij,s} - \frac{3N-2}{N}\mathbb{U}_{1Ns}^*\left(\theta\right) - \frac{3N-2}{N}\mathbb{U}_{2Ns}^*,$$

where $\mathbb{U}_{Ns}^*\left(\theta\right) = \mathbb{U}_{1Ns}^*\left(\theta\right) + 3\mathbb{H}_{2Ns}^*\left(\theta\right) = \binom{N}{2}^{-1}\sum_{1\le i<j\le N} h_s^*(\xi_i^*,\xi_j^*;\theta)$ and $h_s^*(\xi_i^*,\xi_j^*;\theta) = m_{is}^*\left(\theta\right) m_{js}^*\left(\theta\right)\breve{\kappa}_{ij,s}$ with $\breve{\kappa}_{ij,s} = \mathbb{E}_k(\kappa_{ik,s}+\kappa_{jk,s}) - \kappa_{ij,s}$.

Define

$$\begin{aligned}
\mathcal{F}_{1s}^* \equiv \{m_s^*\left(\cdot;\theta_s\right) : (\mathbb{R}^T\times\mathcal{X}^T\times\mathbb{R}) \to \mathbb{R} :\ & m_s^*\left((y,x,v);\theta_s\right) = \{\left[y_s - g\left(x_s\right)\right] + \sum_{r=1}^{R}\phi_{s,r}\left[y_{T-R+r} - g\left(x_{T-R+r}\right)\right]\}v \\
& \text{for some } \theta_s = \left(\phi_s', g\right)' \in \Phi_s\times\mathcal{G}\},
\end{aligned} \tag{A.51}$$

$$\begin{aligned}
\mathcal{F}_{2s}^* \equiv \{f_s^*(\cdot,\cdot;\theta_s) : S^*\times S^* \to \mathbb{R} :\ & f_s^*\left(\xi_1^*,\xi_2^*;\theta_s\right) = m_s^*\left(\xi_1^*;\theta_s\right) m_s^*\left(\xi_2^*;\theta_s\right)\breve{\kappa}_{12,s} \\
& \text{for some } \theta_s = \left(\phi_s', g\right)' \in \Phi_s\times\mathcal{G}\},
\end{aligned} \tag{A.52}$$

[13] Note that Θ_R is a linear subspace, so it is closed, which, in conjunction with the compactness of Θ, implies the compactness of $\Theta \cap \Theta_R$.

and

$$\mathcal{F}_{3s}^{*}\equiv\{f_{s}^{*}(\cdot,\cdot,\cdot;\theta_{s}):S^{*}\times S^{*}\times S^{*}\rightarrow\mathbb{R}:\ f_{s}^{*}\left(\xi_{1}^{*},\xi_{2}^{*},\xi_{3}^{*};\theta_{s}\right)=m_{s}^{*}\left(\xi_{1};\theta_{s}\right)m_{s}^{*}\left(\xi_{2};\theta_{s}\right)\breve{\kappa}_{12,s}$$
$$+m_{s}^{*}\left(\xi_{1};\theta_{s}\right)m_{s}^{*}\left(\xi_{3};\theta_{s}\right)\breve{\kappa}_{13,s}+m_{s}^{*}\left(\xi_{2};\theta_{s}\right)m_{s}^{*}\left(\xi_{3};\theta_{s}\right)\breve{\kappa}_{23,s}$$
$$\text{for some }\theta_{s}=\left(\phi_{s}^{\prime},g\right)^{\prime}\in\Phi_{s}\times\mathcal{G}\},\tag{A.53}$$

where $S^{*}=S\times\mathbb{R}$. It is easy to see the envelope functions for $\mathcal{F}_{1s}^{*}$, $\mathcal{F}_{2s}^{*}$ and $\mathcal{F}_{3s}^{*}$ are respectively given by $F_{1}^{*}\left(\xi^{*}\right)\equiv K\left(|y|+1\right)|v|$, $F_{2}^{*}\left(\xi_{1}^{*},\xi_{2}^{*}\right)=K(|y_{1}|+1)(|y_{2}|+1)\breve{\kappa}_{12,s}\left|v_{1}v_{2}\right|$, and $F_{3}^{*}\left(\xi_{1}^{*},\xi_{2}^{*},\xi_{3}^{*}\right)=K\{(|y_{1}|+1)(|y_{2}|+1)\breve{\kappa}_{12,s}\left|v_{1}v_{2}\right|+(|y_{1}|+1)(|y_{3}|+1)\breve{\kappa}_{13,s}\left|v_{1}v_{3}\right|+(|y_{2}|+1)(|y_{3}|+1)\breve{\kappa}_{23,s}\left|v_{2}v_{3}\right|\}$. Following the analysis in the proof of Theorem 3.1, we can readily show that

$$\log\mathbb{N}_{[]}\left(\epsilon\left\|F_{\ell}^{*}\right\|,\ \mathcal{F}_{\ell s}^{*},\ \left\|\cdot\right\|_{L^{2}}\right)\preceq\ln\left(\frac{1}{\epsilon}\right)+\frac{1}{\epsilon}\text{ for }\ell=1,2,3,\tag{A.54}$$

Part I. Proof of part (i). It is easy to see that $\mathbb{E}^{*}[F_{2}^{*}\left(\xi_{1}^{*},\xi_{2}^{*}\right)^{2}]<\infty$, verifying Condition (a) in Theorem 5.6 of AC. As in the proof of Theorem 3.1, we can readily show that

$$\begin{aligned}
&\lim_{\delta\rightarrow0}\limsup_{N\rightarrow\infty}\mathbb{E}^{o*}\left[\int_{0}^{\delta}\log\mathbb{N}_{N,2}\left(\varepsilon,\mathcal{F}_{2s}^{*}\right)d\epsilon\right]\\
=\;&\lim_{\delta\rightarrow0}\limsup_{N\rightarrow\infty}\mathbb{E}^{o*}\left[\int_{0}^{\delta}\sum_{r=0}^{2}\log\mathbb{N}\left(\frac{\varepsilon}{2\sqrt{3}c_{2,r}},\mathcal{F}_{2s}^{*},\ \left\|\cdot\right\|_{L^{2}\left(U_{N}^{r}\times P^{*2-r}\right)}\right)d\varepsilon\right]\\
\preceq\;&\lim_{\delta\rightarrow0}\limsup_{N\rightarrow\infty}\left[\int_{0}^{\delta}\log\frac{1}{\varepsilon}d\epsilon+\mathbb{E}^{*}\left\{\left[U_{N}^{1}(P^{*1}F_{2}^{2})\right]^{\nu/2}+\left[U_{N}^{2}(F_{2}^{*2})\right]^{\nu/2}\right\}\delta^{1-\nu}\right]=0,
\end{aligned}$$

where $\mathbb{E}^{o*}$ is the outer-expectation associated to $\mathbb{E}^{*}$, the last equality follows from the fact that $\mathbb{E}^{*}\{\left[U_{N}^{2}(F_{2}^{*2})\right]^{\nu/2}\}\leq\left\{\mathbb{E}^{*}\left[U_{N}^{2}(F_{2}^{*2})\right]\right\}^{\nu/2}=\{\mathbb{E}^{*}\left(F_{2}^{*2}\right)\}^{\nu/2}<\infty$ by Jensen inequality and similarly $\mathbb{E}^{*}\{\left[U_{N}^{1}(P^{*1}F_{2}^{*2})\right]^{\nu/2}\}<\infty$. This verifies condition (c) in Theorem 5.6 of AC. Note that $\mathbb{N}\left(\frac{\varepsilon}{2\sqrt{3}c_{2,r}},\mathcal{F}_{2s}^{*},\ \left\|\cdot\right\|_{L^{2}\left(U_{N}^{r}\times P^{*2-r}\right)}\right)=1$ a.s. for $r=0,1,2$ and for sufficiently large ε, say, $\varepsilon\geq\varepsilon_{0}^{*}$. It follows that for some small $\epsilon>0$ and by the above calculations,

$$\begin{aligned}
&\mathbb{E}^{o*}\left|\int_{0}^{\infty}\log\mathbb{N}_{N,2}\left(\varepsilon,\pi_{2,2}\mathcal{F}_{2s}^{*}\right)d\varepsilon\right|^{1+\epsilon}\\
\preceq\;&\left|\int_{0}^{\varepsilon_{0}^{*}}\log\frac{1}{\varepsilon}d\epsilon\right|^{1+\epsilon}+\mathbb{E}\left\{\left[U_{N}^{1}(P^{*1}F_{2}^{*2})\right]^{(1+\epsilon)\nu/2}+\left[U_{N}^{2}(F_{2}^{*2})\right]^{(1+\epsilon)\nu/2}\right\}<\infty,
\end{aligned}$$

where the last inequality holds by choosing ϵ sufficiently small such that $\left(1+\epsilon\right)\nu/2\leq1$. This verifies the uniform integrability of the sequence $\left\{\int_{0}^{\infty}\log N_{N,2}\left(\varepsilon,\mathcal{F}_{2s}^{*}\right)d\varepsilon\right\}_{N=1}^{\infty}$ and thus condition (b) in Theorem 5.6 of AC.

Then by Theorem 5.6 of AC, we have $N\mathbb{U}_{Ns}^{*}\left(\theta\right)\Longrightarrow\mathbb{C}_{s}^{*}\left(\theta\right)$ in $L^{\infty}\left(\Theta\right)$, where $\mathbb{C}_{s}^{*}\left(\theta\right)=\mathbb{C}\left(h_{s}^{*}\left(\cdot,\cdot;\theta\right)\right)$ and $h_{s}^{*}\left(\xi_{i}^{*},\xi_{j}^{*};\theta_{s}\right)=m_{s}^{*}\left(\xi_{i}^{*};\theta_{s}\right)m_{s}^{*}\left(\xi_{j}^{*};\theta_{s}\right)\breve{\kappa}_{12,s}=m_{is}\left(\theta\right)m_{js}\left(\theta\right)v_{i}v_{j}\breve{\kappa}_{12,s}$. We now argue that $\{\mathbb{C}_{s}^{*}\left(\theta\right)\}$ share the same finite-sample distribution as that of $\{\mathbb{C}_{s}\left(\theta\right)=\mathbb{C}\left(h_{s}\left(\cdot,\cdot;\theta\right)\right)\}$ when $\theta\in\Theta_{I}$. That is, $\{N\mathbb{U}_{Ns}^{*}\left(\theta_{(1)}\right),...,N\mathbb{U}_{Ns}^{*}\left(\theta_{(L)}\right)\}$ has the same limiting distribution as $\{N\mathbb{U}_{Ns}\left(\theta_{(1)}\right),...,N\mathbb{U}_{Ns}^{*}\left(\theta_{(L)}\right)\}$ for any finite L when we restrict $\theta_{(1)},...,\theta_{(L)}$ to lie in Θ_{I}. Without loss of generality, we can focus on the case $L=1$.

Note that for $\theta \in \Theta_I$, $h_s\left(\xi_i, \xi_j; \theta\right) = m_{is}(\theta)\, m_{js}(\theta)\, \breve{\kappa}_{12,s}$ and $h_s^*(\xi_i^*, \xi_j^*; \theta_s) = h_s\left(\xi_i, \xi_j; \theta\right) v_i v_j$. Let $\{\lambda_k\}$ denote an enumeration of the positive eigenvalues of $\lambda\Phi(\cdot) = \mathbb{E}\left[h_s(\cdot, \xi_1; \theta)\, \Phi(\xi_1)\right]$ in decreasing order and according to their multiplicity. The corresponding orthonormal eigenfunctions are denoted by $\{\Phi_k(\cdot)\}_{k=1}^{\infty}$. It follows from a version of Mercer's theorem (e.g., Theorem 2 of Sun (2005)) that

$$h_s^{(K)}(\xi, \tilde{\xi}; \theta_s) = \sum_{k=1}^{K} \lambda_k \Phi_k(\xi)\, \Phi_k(\tilde{\xi}) \to \sum_{k=1}^{\infty} \lambda_k \Phi_k(\xi)\, \Phi_k(\tilde{\xi}) \equiv h_s(\xi, \tilde{\xi}; \theta_s)$$

for all ξ and $\tilde{\xi}$ on the support of the probability law of ξ_i. Let

$$\begin{aligned} V_N &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} h_s\left(\xi_i, \xi_j; \theta_s\right) \text{ and } V_N^* = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} h_s\left(\xi_i, \xi_j; \theta_s\right) v_i v_j. \\ V_N^{(K)} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} h_s^{(K)}\left(\xi_i, \xi_j; \theta_s\right) \text{ and } V_N^{(K*)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} h_s^{(K)}\left(\xi_i, \xi_j; \theta_s\right) v_i v_j. \end{aligned}$$

Noting that $V_N - V_N^{(K)} \geq 0$, it is standard to show that

$$\begin{aligned} \mathbb{E}\left|V_N - V_N^{(K)}\right| &= \sum_{k=K+1}^{\infty} \lambda_k \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\Phi_k(\xi_i)^2\right] = \sum_{k=K+1}^{\infty} \lambda_k \to 0 \text{ as } K \to \infty \text{ and} \\ \mathbb{E}^*\left|V_N^* - V_N^{(K*)}\right| &= \sum_{k=K+1}^{\infty} \lambda_k \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}^*\left[\Phi_k(\xi_i)^2 v_i^2\right] = \sum_{k=K+1}^{\infty} \lambda_k \to 0 \text{ as } K \to \infty. \end{aligned}$$

Let $\zeta_i = \left(\Phi_1(\xi_i), ..., \Phi_K(\xi_i)\right)'$ and $\zeta_i^* = \left(\Phi_1(\xi_i), ..., \Phi_K(\xi_i)\right)' v_i$. For any fixed K, it is trivial to show under Assumption 3.1 that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \zeta_i \xrightarrow{\mathcal{L}} N\left(0, \mathbb{I}_K\right) \text{ and } \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \zeta_i^* \xrightarrow{\mathcal{L}} N\left(0, \mathbb{I}_K\right).$$

Then by the continuous mapping theorem, we have

$$\begin{aligned} V_N^{(K)} &= \sum_{k=1}^{K} \lambda_k \left(\frac{1}{N^{1/2}}\sum_{i=1}^{N} \Phi_k(\xi_i)\right)^2 \xrightarrow{\mathcal{L}} \sum_{k=1}^{K} \lambda_k W_k^2 \text{ and} \\ V_N^{(K*)} &= \sum_{k=1}^{K} \lambda_k \left(\frac{1}{N^{1/2}}\sum_{i=1}^{N} \Phi_k(\xi_i)\, v_i\right)^2 \xrightarrow{\mathcal{L}} \sum_{k=1}^{K} \lambda_k W_k^2, \end{aligned}$$

where $\{W_k\}$ is a sequence of independent $N(0,1)$ random variables. These results, in conjunction with Theorem 2 in Dehling, Durieu and Volny (2009), imply that $V_N \xrightarrow{\mathcal{L}} \sum_{k=1}^{\infty} \lambda_k W_k^2$ and $V_N^* \xrightarrow{\mathcal{L}} \sum_{k=1}^{\infty} \lambda_k W_k^2$. Consequently, we have

$$N\mathbb{U}_{Ns}(\theta) = V_N - \frac{1}{N}\sum_{i=1}^{N} h_s(\xi_i, \xi_i; \theta) = V_N - \sum_{k=1}^{\infty} \lambda_k \frac{1}{N}\sum_{i=1}^{N} \Phi_k(\xi_i)^2 \xrightarrow{\mathcal{L}} \sum_{k=1}^{\infty} \lambda_k \left(W_k^2 - 1\right)$$

and

$$N\mathbb{U}_{Ns}^*(\theta) = V_N^* - \frac{1}{N}\sum_{i=1}^{N} h_s^*(\xi_i^*, \xi_i^*; \theta_s) = V_N - \sum_{k=1}^{\infty} \lambda_k \frac{1}{N}\sum_{i=1}^{N} \Phi_k(\xi_i)^2 v_i^2 \xrightarrow{\mathcal{L}} \sum_{k=1}^{\infty} \lambda_k \left(W_k^2 - 1\right).$$

That is, $N\mathbb{U}_{Ns}^*(\theta)$ share the same asymptotic distribution as $N\mathbb{U}_{Ns}(\theta)$ when $\theta \in \Theta_I$. As a result, we have $N\mathbb{U}_{Ns}^*(\theta) \Longrightarrow \mathbb{C}_s(\theta)$ in $L^{\infty}(\Theta_I)$.

Next, note that $\mathbb{H}_{3N}^{*}(\theta)$ is a third order $P^{*}-$canonical $U-$process with the envelope function for its associated kernel given by F_3^{*}. Following the analysis of $\mathbb{U}_N^{*}(\theta)$, it is easy to show that $\mathbb{E}^{*}[F_3^{*}(\xi_1^{*},\xi_2^{*},\xi_3^{*})]<\infty$, $\lim_{\delta\to 0}\limsup_{N\to\infty}\mathbb{E}^{o*}\left[\int_0^{\delta}[\log \mathbb{N}_{N,2}(\varepsilon,\mathcal{F}_{3s}^{*})]^{3/2}d\epsilon\right]=0$ and the sequence $\left\{\int_0^{\infty}[\log N_{N,2}(\varepsilon,\mathcal{F}_{3s}^{*})]^{3/2}d\varepsilon\right\}_{N=1}^{\infty}$ is uniformly integrable. As a result, $N^{3/2}\mathbb{H}_{3N}^{*}(\theta)$ converges to a Gaussian chaos process and $\sup_{\theta\in\Theta}|N\mathbb{H}_{3N}^{*}(\theta)|=O_P\left(N^{-1/2}\right)$. In addition, by the uniform law of large numbers for U-statistics, $\frac{2}{N^2}\sum_{1\le i\neq j\le N}m_{is}^{*}(\theta)^2\kappa_{ij,s}$ $=2\mathbb{E}^{*}\left[\tilde{m}_{1s}(\theta)^2 v_1^2\kappa_{12,s}\right]+o_p(1)=2\mathbb{E}^{*}\left[\tilde{m}_{1s}(\theta)^2\kappa_{12,s}\right]=\mathbb{B}_s(\theta)+o_p(1)$ uniformly in $\theta\in\Theta$. Following the analysis of $\mathbb{U}_{Ns}^{*}(\theta)$, we can also show that both $N\mathbb{U}_{1Ns}^{*}(\theta)$ and $N\mathbb{U}_{2Ns}^{*}(\theta)$ converge to Gaussian chaos processes. Consequently, we have

$$S_{Ns}^{*}(\theta)\Longrightarrow\mathbb{B}_s(\theta)+\mathbb{C}_s^{*}(\theta), \tag{A.55}$$

where the process $\{\mathbb{C}_s^{*}(\theta)\}$ coincides with $\{\mathbb{C}_s(\theta)\}$ on Θ_I.

Part II. Proof of part (ii). When $\theta\notin\Theta_I$, (A.55) continues to hold, which implies that

$$\mu_N^{-1}S_N^{*}(\theta)=\mu_N^{-1}\sum_{s=1}^{T-R}S_{Ns}^{*}(\theta)=O_P\left(\mu_N^{-1}\right)\ \text{uniformly in }\theta\in\Theta\backslash\Theta_I.$$

By Theorem 3.1(ii) and Assumption , we can show that

$$\min_{\theta\in\Theta_N\cap\Theta_R}\frac{1}{N}S_N(\theta)\xrightarrow{p}\min_{\theta\in\Theta\cap\Theta_R}\sum_{s=1}^{T-R}\{\text{MDD}[m_s(Y,X,\theta)|\underline{z}_s]\}^2.$$

Then under $\mathbb{H}_1:\Theta_I\cap\Theta_R=\emptyset$,

$$\begin{aligned}\mu_N^{-1}\hat{S}_N^{*} &= \min_{\theta\in\Theta_N\cap\Theta_R}\left[\mu_N^{-1}S_N^{*}(\theta)+\frac{1}{N}S_N(\theta)\right]=\min_{\theta\in\Theta_N\cap\Theta_R}\left[\frac{1}{N}S_N(\theta)\right]+O_p\left(\mu_N^{-1}\right)\\ &\xrightarrow{p}\min_{\theta\in\Theta\cap\Theta_R}\sum_{s=1}^{T-R}\{\text{MDD}[m_s(Y,X,\theta)|\underline{z}_s]\}^2.\end{aligned}$$

This completes the proof of the theorem. ■

REFERENCES

Aguiar, M., Bils, M., 2015. Has consumption inequality mirrored income inequality? American Economic Review 105 (9), 2725–2756.

Andrews, D. W., Shi, X., 2013. Inference based on conditional moment inequalities. Econometrica 81, 609–666.

Arcones, M. A., Giné, E., 1993. Limit theorems for U-processes. Annals of Probability 21, 1494–1542.

Ahn, S., Horenstein, A., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.

Ahn, S. C., Lee, Y.H, Schmidt, P., 2001. GMM estimation of linear panel data models with time-varying individual effects. Journal of Econometrics 101, 219-255.

Ahn, S. C., Lee, Y.H., Schmidt, P., 2013. Panel data models with multiple time-varying individual effects, Journal of Econometrics 174, 1–14.

Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, 1795–1843.

Andrews, D. W., 1994. Empirical process methods in econometrics. In Handbook of Econometrics, vol. 4, R.F. Engle and D. L. McFadden (eds), pp. 2248–2294.

Andrews, D. W., Shi, X., 2014. Nonparametric inference based on conditional moment inequalities. Journal of Econometrics 179, 31–45.

Andrews, D. W., Shi, X., 2017. Inference based on many conditional moment inequalities. Journal of Econometrics 196, 275–287.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.

Banks, J., Blundell, R., Lewbel, A. 1997. Quadratic Engel curves and consumer demand. Review of Economics and Statistics 79, 527–539.

Bierens, H., 1982. Consistent model specification tests. Journal of Econometrics 20, 105–134.

Blundell, R. W., Browning, M., Crawford, I. A., 2003. Nonparametric Engel curves and revealed preference. Econometrica 71 (1), 205–240.

Blundell, R. W., Chen, X., Kristensen, D., 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. Econometrica 75 (6), 1613–1669.

Browning, M., Crossly, T., 2009. Are two cheap, noisy measures better than one expensive, accurate one? American Economic Review 99 (2), 99–103.

Chen, X., Pouzo, D., 2012. Estimation of nonparametric conditional moment models with possibly non-smooth moments. Econometrica 80, 277–321.

Coakley, J., Fuertes, A., Smith, R., 2002. A principal components approach to cross-section dependence in panels. Working paper, Birkbeck College, University of London..

Chernozhukov, V., Newey, W., Santos, A., 2015. Constrained conditional moment restriction models. Working paper, Dept. of Economics, MIT.

Deaton, A. S., Muellbauer, J., 1980. An almost ideal demand system. American Economic Review 70 (3), 312–326.

Dehling, H., Durieu, O., Volny, D., 2009. New techniques for empirical processes of dependent data. Stochastic Processes and their Applications 119, 3699–3718.

de la, Peña, V. H., Giné, E., 1999. Decoupling: From Dependence to Independence. Springer, New York.

Dominguez, M. A., Lobato, I. N., 2004. Consistent estimation of models defined by conditional moment restrictions. Econometrica 72, 1601–1615.

Dong, C., Gao, J., Peng, B., 2018. Varying–coefficient panel data models with nonstationarity and partially observed factor structure. Working Paper, Monash University.

Freyberger, J., 2018. Non-parametric panel data models with interactive fixed effects. Review of Economic Studies, forthcoming.

Greenaway-McGrevy, R., Han, C., Sul, D., 2008. Estimating and testing idiosyncratic equations using cross-section dependent panel data. Working Paper, Dept. of Economics, Univ. of Auckland..

Hamilton, B. W., 2001. Using Engel's law to estimate CPI bias. American Economic Review 91 (3), 619–630.

Holz-Eakin, D., Newey, W., Rosen, H., 1988. Estimating vector autoregressions with panel data. Econometrica 56, 1371–1395.

Hong, S., 2017. Inference in semiparametric conditional moment models with partial identification. Journal of Econometrics 196, 156–179.

Hurst, E., Li, G., Pugsley, B., 2014. Are household surveys like tax forms? evidence from income underreporting of the self-employed. Review of Economics and Statistics 96 (1), 19–33.

Kapetanios, G., Pesaran, M. H., 2007. Alternative approaches to estimation and inference in large multifactor panels: small sample results with an application to modelling of asset returns. The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis, Cambridge University Press, Cambridge.

Kororok, M. R., 2008. Introduction to Empirical Processes and Semiparametric Inference. Springer, New York.

Lee, A. J., 1990. U-statistics: Theory and Practice. CRC Press, New York.

Leucht, A., Neumann, M. H., 2013. Dependent wild bootstrap for degenerate U- and V-statistics. Journal of Multivariate Analysis 117, 257–280.

Lu, X., Su, L., 2016. Shrinkage estimation of dynamic panel data models with interactive fixed effects. Journal of Econometrics 190, 148–175.

Manski, C. F., 2003. Partial identification of probability distributions. Springer-Verlag, New York.

Moon, H. R., Weidner, M., 2015. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica 83, 1543–1579.

Moon, H. R., Weidner, M., 2017. Dynamic linear panel regression models with interactive fixed effects. Econometric Theory 33, 158-195.

Nakamura, E., Steinsson, J., Liu, M., 2016. Are Chinese growth and inflation too smooth? evidence from Engel curves. American Economic Journal: Macroeconomics 8 (3), 113–144.

Newey, W. K., Powell, J. L., 2003. Instrumental variable estimation of nonparametric models. Econometrica 71, 1565–1578.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Review of Economics and Statistics 92, 1004–1016.

Pesaran, M. H., 2006. Estimation and inference in large heterogenous panels with multifactor error. Econometrica 74, 967–1012.

Pesaran, M. H., Tosetti, E., 2007. Large Panels with Common Factors and Spatial Correlation. Journal of Econometrics 161, 182–202.

Phillips, P. C. B., Sul, D., 2003. Dynamic panel estimation and homogeneity testing under cross sectional dependence. Econometrics Journal 6, 217–259.

Phillips, P. C. B., Sul, D., 2007. Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. Journal of Econometrics 137, 162–188.

Pissarides, C. A., Weber, G., 1989. An expenditure-based estimate of Britain's black economy. Journal of Public Economics 39, 17–32.

Santos, A., 2012. Inference in nonparametric instrumental variables with partial identification. Econometrica 80, 213–275.

Schumaker, L., 2007. Spline Functions: Basic Theory. Cambridge University Press, Cambridge.

Shao, X., Zhang, J., 2014. Martingale divergence correlation and its use in high dimensional variable screening. Journal of the American Statistical Association 109, 1302-1318.

Stinchcombe, M. B., White, H., 1998. Consistent specification testing with nuisance parameters present only under the alternative. Econometric Theory 14, 295-325.

Su, L., Jin, S., 2012. Sieve estimation of panel data models with cross section dependence. Journal of Econometrics 169, 34–47.

Su, L., Jin, S., Zhang, Y., 2015. Specification test for panel data models with interactive fixed effects. Journal of Econometrics 186, 222–244.

Su, L., Miao, K., Jin, S., 2019. On factor models with random missing: EM estimation, inference, and cross validation. Working Paper, Singapore Management University.

Su, L., Zhang, Y., 2018. Nonparametric dynamic panel data models with interactive fixed effects: sieve estimation and specification testing. Working Paper, Singapore Management University.

Su, L., Zheng, X., 2017. A Martingale-difference-divergence-based test for specification. Economics Letters 156, 162–167.

Sun, H., 2005. Mercer theorem for RKHS on noncompact sets. Journal of Complexity 21, 337–349.

van der Vaart, A. W., Wellner, J., 1996. Weak Convergence and Empirical Processes: with applications to statistics. Springer, New York.

Online Supplement for
"Inference in Partially Identified Panel Data Models with Interactive Fixed Effects"

Shengjie Hong[a], Liangjun Su[b], Yaqi Wang[c]

[a] School of Economics and Management, Tsinghua University

[b] School of Economics, Singapore Management University

[c] School of Finance, Central University of Finance and Economics

This online supplement contains the proofs of the technical lemmas in Appendix A, as well as a discussion on the sufficient conditions for Assumption 3.4 to hold.

B Proofs of Lemmas A.1–A.6

Proof of Lemma A.1. The proof is mainly based on the definitions of MDD_o and MDD as specified in Shao and Zhang (2014) and Su and Zheng (2017), respectively, as well as their properties as shown in these two papers. Let d_Z denote the dimension of Z.

We first prove the first claim. Note that $\mathrm{MDD}_o\left(W_1|Z\right)^2=0$ if and only if $\mathbb{E}\left(W_1|Z\right)=\mathbb{E}\left(W_1\right)$, which, in turn, implies that for any given constant vector $s\in\mathbb{R}^{d_Z}$,

$$\begin{aligned}\mathrm{Cov}\left(W_1,\ \exp\left(\mathbf{i}s'Z\right)\right) &= \mathbb{E}\left[W_1\exp\left(\mathbf{i}s'Z\right)\right]-\mathbb{E}\left[W_1\right]\mathbb{E}\left[\exp\left(\mathbf{i}s'Z\right)\right]\\ &= \mathbb{E}\left[\mathbb{E}\left(W_1|Z\right)\exp\left(\mathbf{i}s'Z\right)\right]-\mathbb{E}\left[W_1\right]\mathbb{E}\left[\exp\left(\mathbf{i}s'Z\right)\right]\\ &= \mathbb{E}\left[W_1\right]\mathbb{E}\left[\exp\left(\mathbf{i}s'Z\right)\right]-\mathbb{E}\left[W_1\right]\mathbb{E}\left[\exp\left(\mathbf{i}s'Z\right)\right]\\ &= 0,\end{aligned}$$

where the second equality follows from the law of iterated expectations. Then by equation (2.4) in Su and Zheng (2017), we have

$$\begin{aligned}\mathrm{MDD}_o\left(W_2-W_1|Z\right)^2 &= \int_{\mathbb{R}^{d_Z}}\left[\mathrm{Cov}\left(W_2-W_1,\exp\left(\mathbf{i}s'Z\right)\right)\right]^2\cdot q\left(s\right)ds\\ &= \int_{\mathbb{R}^{d_Z}}\left[\mathrm{Cov}\left(W_2,\exp\left(\mathbf{i}s'Z\right)\right)-\mathrm{Cov}\left(W_1,\exp\left(\mathbf{i}s'Z\right)\right)\right]^2\cdot q\left(s\right)ds\\ &= \int_{\mathbb{R}^{d_Z}}\left[\mathrm{Cov}\left(W_2,\exp\left(\mathbf{i}s'Z\right)\right)\right]^2\cdot q\left(s\right)ds\\ &= \mathrm{MDD}_o\left(W_2|Z\right)^2,\end{aligned}\tag{B.1}$$

where $\mathbf{i}\equiv\sqrt{-1}$, $q\left(s\right)\equiv 1/\left[c\left|s\right|^{(1+d_Z)}\right]$, $c\equiv\pi^{(1+d_Z)/2}/\Gamma\left(\frac{1+d_Z}{2}\right)$, and $\Gamma\left(\cdot\right)$ is the complete gamma function: $\Gamma\left(z\right)\equiv\int_0^\infty t^{(z-1)}\exp\left(-t\right)dt$.

To proceed, note that for a generic real-valued random variable W,

$$\mathrm{MDD}\left(W|Z\right)^2=\mathrm{MDD}_o\left(W|Z\right)^2+[\mathbb{E}\left(W\right)]^2\mathbb{E}\left(\left|Z-Z^\dagger\right|\right)\tag{B.2}$$

which follows directly from the definitions of MDD_o and MDD. Also note that $\mathrm{MDD}\left(W_1|Z\right)^2 = 0$ if and only if $\mathbb{E}\left(W_1|Z\right) = 0$. It follows that $\mathrm{MDD}\left(W_1|Z\right)^2 = 0$ if and only if

$$\mathrm{MDD}_o\left(W_1|Z\right)^2 = 0 \text{ and } \mathbb{E}\left(W_1\right) = 0. \tag{B.3}$$

This, in conjunction with the given condition that $\mathrm{MDD}\left(W_1|Z\right)^2 = 0$, implies that

$$\begin{aligned}
\mathrm{MDD}\left(W_2 - W_1|Z\right)^2 &= \mathrm{MDD}_o\left(W_2 - W_1|Z\right)^2 + \left[\mathbb{E}\left(W_2 - W_1\right)\right]^2 \mathbb{E}\left(\left|Z - Z^\dagger\right|\right) \\
&= \mathrm{MDD}_o\left(W_2 - W_1|Z\right)^2 + \left[\mathbb{E}\left(W_2\right)\right]^2 \mathbb{E}\left(\left|Z - Z^\dagger\right|\right) \\
&= \mathrm{MDD}_o\left(W_2|Z\right)^2 + \left[\mathbb{E}\left(W_2\right)\right]^2 \mathbb{E}\left(\left|Z - Z^\dagger\right|\right) \\
&= \mathrm{MDD}\left(W_2|Z\right)^2,
\end{aligned}$$

where the first and last equalities follow from(B.2), the second equality holds by (B.3), and the third equality holds by (B.1). This proves the first claim.

Now, we prove the second claim. By Su and Zheng (2017),

$$\begin{aligned}
\mathrm{MDD}\left(W_2 - W_1|Z\right)^2 &= -\mathbb{E}\left[\left(W_2 - W_1\right)\left(W_2^\dagger - W_1^\dagger\right)\left|Z - Z^\dagger\right|\right] \\
&\quad + 2\mathbb{E}\left[\left(W_2 - W_1\right)\left|Z - Z^\dagger\right|\right]\mathbb{E}\left[W_2^\dagger - W_1^\dagger\right] \\
&= -\mathbb{E}\left[\left(W_2 - W_1\right)\left(W_2^\dagger - W_1^\dagger\right)\left|Z - Z^\dagger\right|\right] \\
&\quad + 2\mathbb{E}\left[W_2\left|Z - Z^\dagger\right|\right]\mathbb{E}\left[W_2^\dagger\right] - 2\mathbb{E}\left[W_1\left|Z - Z^\dagger\right|\right]\mathbb{E}\left[W_2^\dagger - W_1^\dagger\right] \\
&\quad - 2\mathbb{E}\left[W_2\left|Z - Z^\dagger\right|\right]\mathbb{E}\left[W_1^\dagger\right].
\end{aligned}$$

Noting that $\mathbb{E}\left(W_1|Z\right) = 0$, we have $\mathbb{E}\left(W_1\right) = \mathbb{E}(W_1^\dagger) = 0$ and $\mathbb{E}\left[W_1\left|Z - Z^\dagger\right|\right] = 0$ by the law of iterated expectations and the independence between (W_1, Z) and $Z^\dagger$. Then we have

$$\mathrm{MDD}\left(W_2 - W_1|Z\right)^2 = -\mathbb{E}\left[\left(W_2 - W_1\right)\left(W_2^\dagger - W_1^\dagger\right)\left|Z - Z^\dagger\right|\right] + 2\mathbb{E}\left[W_2\left|Z - Z^\dagger\right|\right]\mathbb{E}\left[W_2^\dagger\right]. \tag{B.4}$$

And we also know that

$$\mathrm{MDD}\left(W_2|Z\right)^2 = -\mathbb{E}\left[W_2 W_2^\dagger\left|Z - Z^\dagger\right|\right] + 2\mathbb{E}\left[W_2\left|Z - Z^\dagger\right|\right]\mathbb{E}\left[W_2^\dagger\right]. \tag{B.5}$$

(B.4)–(B.5), in conjunction with the fact that $\mathrm{MDD}\left(W_2 - W_1|Z\right)^2 = \mathrm{MDD}\left(W_2|Z\right)^2$, implies that

$$\mathbb{E}\left[\left(W_2 - W_1\right)\left(W_2^\dagger - W_1^\dagger\right)\left|Z - Z^\dagger\right|\right] = \mathbb{E}\left[W_2 W_2^\dagger\left|Z - Z^\dagger\right|\right]. \blacksquare$$

Proof of Lemma A.2. Note that for any real-valued random variable W, it holds that

$$\mathrm{MDD}\left(W|Z\right)^2 = \mathrm{MDD}_o\left(W|Z\right)^2 + \left[\mathbb{E}\left(W\right)\right]^2\mathbb{E}\left(\left|Z - Z^\dagger\right|\right). \tag{B.6}$$

We prove the first and second inequalities in Parts I and II, respectively.

Part I. We organize Part I into three subparts. In Part I (i), we show the existence of a finite constant b_1 s.t. for any pair $\{W_1, W_2\} \subseteq \mathcal{W}$, it holds that

$$\left|\mathrm{MDD}_o\left(W_1|Z\right)^2-\mathrm{MDD}_o\left(W_2|Z\right)^2\right| \leq b_1 \cdot\left[\mathrm{MDD}_o\left(W_1-W_2|Z\right)^2\right]^{1/2}.$$

In Part I (ii), we show the existence of a finite constant b_2 s.t. for any pair $\{W_1, W_2\} \subseteq \mathcal{W}$, it holds that

$$\left|\mathbb{E}\left(W_1\right)^2 \mathbb{E}\left(\left|Z-Z^{\dagger}\right|\right)-\mathbb{E}\left(W_2\right)^2 \mathbb{E}\left(\left|Z-Z^{\dagger}\right|\right)\right| \leq b_2 \cdot\left[\mathbb{E}\left(W_1-W_2\right)^2 \mathbb{E}\left(\left|Z-Z^{\dagger}\right|\right)\right]^{1/2}.$$

And in Part I (iii), we combine the results from Part I (i) and Part I (ii) via B.6 to prove the fist inequality.

Part I (i). By the definition of $\mathcal{W}$,

$$\sup_{W \in \mathcal{W}} \operatorname{Var}(W)=\sup_{W \in \mathcal{W}}\left[\mathbb{E}\left(W^2\right)-\mathbb{E}(W)^2\right] \leq \sup_{W \in \mathcal{W}} \mathbb{E}\left(W^2\right) \equiv b_3<\infty. \tag{B.7}$$

Denote by $\varphi_Z(s) \equiv \mathbb{E}\left[\exp \left(\mathbf{i} s^{\prime} Z\right)\right]$, the characteristic function of Z. It holds that

$$\begin{aligned}
\left|\operatorname{Var}\left(\exp \left(\mathbf{i} s^{\prime} Z\right)\right)\right| & =\left|\mathbb{E}\left[\exp \left(\mathbf{i} s^{\prime} Z\right)^2\right]-\mathbb{E}\left[\exp \left(\mathbf{i} s^{\prime} Z\right)\right]^2\right| \\
& =\left|\varphi_Z(2 s)-\left[\varphi_Z(s)\right]^2\right| \leq\left|\varphi_Z(2 s)\right|+\left|\varphi_Z(s)\right|^2 \leq 2,
\end{aligned} \tag{B.8}$$

where the last inequality follows from the fact that $\left|\varphi_Z(\cdot)\right| \leq 1$. By equation (2.4) in Su and Zheng (2017), we have that for any pair $\{W_1, W_2\} \subseteq \mathcal{W}$,

$$\begin{aligned}
& \left|\mathrm{MDD}_o\left(W_1|Z\right)^2-\mathrm{MDD}_o\left(W_2|Z\right)^2\right| \\
= & \left|\int_{\mathbb{R}^{d_Z}}\left[\operatorname{Cov}\left(W_1, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)^2-\operatorname{Cov}\left(W_2, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)^2\right] q(s) d s\right| \\
\leq & \int_{\mathbb{R}^{d_Z}}\left|\operatorname{Cov}\left(W_1, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)^2-\operatorname{Cov}\left(W_2, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)^2\right| q(s) d s \\
\leq & \int_{\mathbb{R}^{d_Z}}\left|\operatorname{Cov}\left(W_1, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)-\operatorname{Cov}\left(W_2, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)\right| \\
& \quad \times\left[\left|\operatorname{Cov}\left(W_1, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)\right|+\left|\operatorname{Cov}\left(W_2, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)\right|\right] q(s) d s \\
\leq & \int_{\mathbb{R}^{d_Z}}\left|\operatorname{Cov}\left(W_1, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)-\operatorname{Cov}\left(W_2, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)\right| \\
& \quad \times\left[\left|\operatorname{Var}\left(W_1\right)^{1/2} \operatorname{Var}\left(\exp \left(\mathbf{i} s^{\prime} Z\right)\right)^{1/2}\right|+\left|\operatorname{Var}\left(W_2\right)^{1/2} \operatorname{Var}\left(\exp \left(\mathbf{i} s^{\prime} Z\right)\right)^{1/2}\right|\right] q(s) d s \\
\leq & 2 \sqrt{2 b_3} \int_{\mathbb{R}^{d_Z}}\left|\operatorname{Cov}\left(W_1, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)-\operatorname{Cov}\left(W_2, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)\right| q(s) d s \\
= & 2 \sqrt{2 b_3} \int_{\mathbb{R}^{d_Z}}\left|\operatorname{Cov}\left(W_1-W_2, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)\right| q(s) d s \\
\leq & 2 \sqrt{2 b_3}\left[\int_{\mathbb{R}^{d_Z}}\left|\operatorname{Cov}\left(W_1-W_2, \exp \left(\mathbf{i} s^{\prime} Z\right)\right)\right|^2 q(s) d s\right]^{1/2} \\
= & 2 \sqrt{2 b_3}\left[\mathrm{MDD}_o\left(W_1-W_2|Z\right)^2\right]^{1/2}
\end{aligned}$$

where $\mathbf{i} \equiv \sqrt{-1}$, $q(s)$ is as defined in the proof of Lemma A.1, the fourth inequality follows from (B.7)-(B.8), and the last inequality follows from the Hölder inequality. Consequently, we have that for any pair $\{W_1, W_2\} \subseteq \mathcal{W}$,

$$\left|\mathrm{MDD}_o\left(W_1|Z\right)^2-\mathrm{MDD}_o\left(W_2|Z\right)^2\right| \leq b_1\left[\mathrm{MDD}_o\left(W_1-W_2|Z\right)^2\right]^{1/2} \tag{B.9}$$

where $b_1 \equiv 2\sqrt{2b_3}$.

Part I (ii). By Jensen inequality and (B.7), we have

$$\sup_{W\in\mathcal{W}} \mathbb{E}\,|W| \leq \sup_{W\in\mathcal{W}} \left[\mathbb{E}\left(W^2\right)\right]^{1/2} \;=\; \sqrt{b_3}. \tag{B.10}$$

For any pair $\{W_1, W_2\} \subseteq \mathcal{W}$,

$$\begin{aligned}
& \left|[\mathbb{E}\,(W_1)]^2\,\mathbb{E}\left|Z - Z^\dagger\right| - [\mathbb{E}\,(W_2)]^2\,\mathbb{E}\left|Z - Z^\dagger\right|\right| \\
= \;& \left|\mathbb{E}\,(W_1 + W_2)\,\mathbb{E}\left|Z - Z^\dagger\right|^{1/2}\right| \left|\mathbb{E}\,(W_1 - W_2)\left(\mathbb{E}\left|Z - Z^\dagger\right|\right)^{1/2}\right| \\
\leq \;& [\mathbb{E}\,|W_1| + \mathbb{E}\,|W_2|]\left(\mathbb{E}\left|Z - Z^\dagger\right|\right)^{1/2}\left[[\mathbb{E}\,(W_1 - W_2)]^2\,\mathbb{E}\left|Z - Z^\dagger\right|\right]^{1/2} \\
\leq \;& 2\sqrt{b_3}\left(\mathbb{E}\left|Z - Z^\dagger\right|\right)^{1/2}\left[[\mathbb{E}\,(W_1 - W_2)]^2\,\mathbb{E}\left|Z - Z^\dagger\right|\right]^{1/2} \\
= \;& b_2\left[[\mathbb{E}\,(W_1 - W_2)]^2\,\mathbb{E}\left|Z - Z^\dagger\right|\right]^{1/2}
\end{aligned} \tag{B.11}$$

where the last inequality follows from (B.9), and $b_2 \equiv 2\sqrt{b_3}\cdot\mathbb{E}\left(\left|Z - Z^\dagger\right|\right)^{1/2} < \infty$ by the condition that $\mathbb{E}\left|Z - Z^\dagger\right| < \infty$ and the fact that $b_3 < \infty$.

Part I (iii). For any pair $\{W_1, W_2\} \subseteq \mathcal{W}$, it holds that

$$\begin{aligned}
& \left|\mathrm{MDD}\,(W_1|Z)^2 - \mathrm{MDD}\,(W_2|Z)^2\right| \\
\leq \;& \left|\mathrm{MDD}_o\,(W_1|Z)^2 - \mathrm{MDD}_o\,(W_2|Z)^2\right| + \left|\left\{[\mathbb{E}\,(W_1)]^2 - [\mathbb{E}\,(W_2)]^2\right\}\mathbb{E}\left(\left|Z - Z^\dagger\right|\right)\right| \\
\leq \;& b_1\left[\mathrm{MDD}_o\,(W_1 - W_2|Z)^2\right]^{1/2} + b_2\left[[\mathbb{E}\,(W_1 - W_2)]^2\,\mathbb{E}\left|Z - Z^\dagger\right|\right]^{1/2} \\
\leq \;& \max\{b_1, b_2\}\left\{\left[\mathrm{MDD}_o\,(W_1 - W_2|Z)^2\right]^{1/2} + \left[[\mathbb{E}\,(W_1 - W_2)]^2\,\mathbb{E}\left|Z - Z^\dagger\right|\right]^{1/2}\right\} \\
\leq \;& \sqrt{2}\max\{b_1, b_2\}\left\{\mathrm{MDD}_o\,(W_1 - W_2|Z)^2 + [\mathbb{E}\,(W_1 - W_2)]^2\,\mathbb{E}\left|Z - Z^\dagger\right|\right\}^{1/2} \\
= \;& b\left[\mathrm{MDD}\,(W_1 - W_2|Z)^2\right]^{1/2},
\end{aligned}$$

where $b \equiv \sqrt{2}\max\{b_1, b_2\} < \infty$, the first inequality follows from the triangle inequality, and the second inequality holds by (B.9) and (B.11).

Part II. It is easy to see that

$$\mathbb{E}\,(W_1 - W_2)^2\,\mathbb{E}\left|Z - Z^\dagger\right| \leq 2\left[\mathbb{E}\,(W_1)^2\,\mathbb{E}\left|Z - Z^\dagger\right| + \mathbb{E}\,(W_2)^2\,\mathbb{E}\left|Z - Z^\dagger\right|\right]. \tag{B.12}$$

Next, note that

$$\begin{aligned}
& \mathrm{MDD}_o\,(W_1 - W_2|Z)^2 \\
= \;& \int_{\mathbb{R}^{d_Z}} \mathrm{Cov}\,(W_1 - W_2, \exp\,(\mathbf{i}s'Z))^2\,q\,(s)\,ds \\
= \;& \int_{\mathbb{R}^{d_Z}} [\mathrm{Cov}\,(W_1, \exp\,(\mathbf{i}s'Z)) - \mathrm{Cov}\,(W_2, \exp\,(\mathbf{i}s'Z))]^2\,q\,(s)\,ds \\
\leq \;& 2\int_{\mathbb{R}^{d_Z}} \mathrm{Cov}\,(W_1, \exp\,(\mathbf{i}s'Z))^2\,q\,(s)\,ds + 2\int_{\mathbb{R}^{d_Z}} \mathrm{Cov}\,(W_2, \exp\,(\mathbf{i}s'Z))^2\,q\,(s)\,ds \\
= \;& 2\left[\mathrm{MDD}_o\,(W_1|Z)^2 + \mathrm{MDD}_o\,(W_2|Z)^2\right].
\end{aligned} \tag{B.13}$$

Combining (B.6), (B.12) and (B.13) yields $\mathrm{MDD}(W_1 - W_2|Z)^2 \leq 2\left[\mathrm{MDD}(W_1|Z)^2 + \mathrm{MDD}(W_2|Z)^2\right]$. It follows that

$$\begin{aligned}\left[\mathrm{MDD}(W_1 - W_2|Z)^2\right]^{1/2} &\leq \sqrt{2}\left[\mathrm{MDD}(W_1|Z)^2 + \mathrm{MDD}(W_2|Z)^2\right]^{1/2} \\ &\leq 2\left\{\left[\mathrm{MDD}(W_1|Z)^2\right]^{1/2} + \left[\mathrm{MDD}(W_2|Z)^2\right]^{1/2}\right\}\end{aligned}$$

where the second inequality follows from the fact that $\left(a^2+b^2\right)^{1/2} \leq \sqrt{2}(a+b)$ for any $a, b \geq 0$. This completes the proof. ■

Proof of Lemma A.3. To prove the first claim, note that $\left(\mathbb{R}^{(T-R)R} \times \mathcal{W}^s(\mathcal{X}), \|\cdot\|_c\right)$ forms a metric space, in which compactness is equivalent to sequential compactness. So it is sufficient to show that $(\Theta, \|\cdot\|_c)$ is sequentially compact.

Recall that $\Theta = \mathbf{\Phi} \times \mathcal{G}$, where $\mathbf{\Phi}$ is compact by assumption. By Lemma A.2 in Santos (2012), $(\mathcal{G}, \|\cdot\|_c)$ is also compact despite the difference in notations. Specifically, $\lfloor \frac{d}{2} \rfloor$ is equivalent to Santos's (2012) m, and $d - \lfloor \frac{d}{2} \rfloor$ is equivalent to his m_0. In addition, the condition $d \geq d_x + 2$ required by Assumption 2.1(ii) guarantees '$\min\{m_0, m\} > \frac{d_x}{2}$', which is required in Santos (2012). Consequently, both $\mathbf{\Phi}$ and $\mathcal{G}$ are sequentially compact. Then for any given sequence $\{\theta_N = \left(\phi_N', g_N\right)'\}$ in Θ, there is subsequence $\{\theta_{M_N} = \left(\phi_{M_N}', g_{M_N}\right)'\}$ s.t. $\phi_{M_N} \to \phi$ under $|\cdot|$ for some $\phi \in \mathbf{\Phi}$ as $M_N \to \infty$, and there is also a subsubsequence $\{\theta_{L_{M_N}} = (\phi_{L_{M_N}}', g_{L_{M_N}})'\}$ in $\{\theta_{M_N} = \left(\phi_{M_N}', g_{M_N}\right)'\}$ s.t. $g_{L_{M_N}} \to g$ under $\|\cdot\|_c$ for some $g \in \mathcal{G}$ as $L_{M_N} \to \infty$. In short, for any given sequence $\{\theta_N = \left(\phi_N', g_N\right)'\}$ in Θ, we are able to find a subsequence $\{\theta_{L_{M_N}} = (\phi_{L_{M_N}}', g_{L_{M_N}})'\}$ s.t. $\theta_{L_{M_N}} \to \left(\phi', g\right)' \in \Theta$ under $\|\cdot\|_c$ as $L_{M_N} \to \infty$. This shows that $(\Theta, \|\cdot\|_c)$ is sequentially compact.

To prove the second claim, note that $(\mathcal{W}^s(\mathcal{X}), \|\cdot\|_c)$ is a metric space. The compactness of its subset $(\mathcal{G}, \|\cdot\|_c)$ as proven above implies total boundedness of $\mathcal{G}$ under $\|\cdot\|_c$, which in turn implies boundedness of $\mathcal{G}$ under $\|\cdot\|_c$. So there exists a constant B_c s.t. $\sup_{g \in \mathcal{G}} \|g\|_c \leq B_c$. As a result, for any given vector of nonnegative integer λ with $\langle\lambda\rangle \leq \frac{d}{2}$, it hods that

$$\begin{aligned}\sup_{x \in \mathcal{X}} \left|D^\lambda g(x)\right| &\leq \sup_{x \in \mathcal{X}} \left|D^\lambda g(x)\right| (1 + x'x)^{\zeta/2} \leq \max_{\langle\lambda\rangle \leq \frac{d}{2}} \left[\sup_{x \in \mathcal{X}} \left|D^\lambda g(x)\right| (1 + x'x)^{\zeta/2}\right] \\ &= \|g\|_c \leq \sup_{g \in \mathcal{G}} \|g\|_c \leq B_c,\end{aligned}$$

where the first inequality follows from the fact that $(1 + x'x)^{\zeta/2} \geq 1$ under the requirement that $\zeta > \left(\frac{d_x}{2} \cdot \lfloor \frac{d}{2} \rfloor\right) / \left(\lfloor \frac{d}{2} \rfloor - \frac{d_x}{2}\right) > 0$ as specified in Assumption 2.1(ii). ■

Proof of Lemma A.4. Here we follow the same notations for various bounds as we have adopted in the previous proofs. Specifically, the compactness of $\mathbf{\Phi}$ according to Assumption 2.1(i) implies that $B_{\mathbf{\Phi}} \equiv \sup_{\phi \in \mathbf{\Phi}} |\phi| < \infty$. By Lemma A.3, for all $g \in \mathcal{G}$, it holds that $\sup_{x \in \mathcal{X}} |g(x)| \leq B_c < \infty$. Then for any $\theta \in \Theta$

and $s = 1, ..., T - R$, we have

$$\begin{aligned}
|m_s(Y, X, \theta)| &= \left| [y_s - g(x_s)] + \sum_{r=1}^{R} \phi_{s,r} [y_{T-R+r} - g(x_{T-R+r})] \right| \\
&\leq [|y_s| + |g(x_s)|] + \sum_{r=1}^{R} |\phi_{s,r}| [|y_{T-R+r}| + |g(x_{T-R+r})|] \\
&\leq |Y| + B_c + B_{\Phi} R [|Y| + B_c] = (B_{\Phi} R + 1) [|Y| + B_c]
\end{aligned}$$

where $Y = (y_1, ..., y_T)'$, $X = (x_1, ..., x_T)'$, $\phi_{s,r}$ denote the rth element in ϕ_s, and the first inequality holds by the triangle inequality. It follows that for any $\theta \in \Theta$ and $s = 1, ..., T - R$,

$$\mathbb{E}\left\{[m_s(Y, X, \theta)]^2\right\} \leq 2(B_{\Phi} R + 1)^2 \left[\mathbb{E}(|Y|^2) + B_c^2\right] < \infty$$

where the last inequality follows from the fact that $\mathbb{E}(|Y|^2) < \infty$ under Assumption 3.1(ii). In addition, by the triangle inequality, Jensen inequality, and Assumption 3.1(ii), $\mathbb{E}\left|\underline{z}_s - \underline{z}_s^{\dagger}\right| \leq \mathbb{E}\left|Z - Z^{\dagger}\right| \leq 2\left[\mathbb{E}(|Z|^2)\right]^{1/2} < \infty$. Therefore, the result in Lemma A.2 is applicable here.

By Lemma A.2, we have that for any $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$,

$$\begin{aligned}
\left|\mathrm{MDD}[m_s(Y, X, \theta_1) \mid \underline{z}_s]^2 - \mathrm{MDD}[m_s(Y, X, \theta_2) \mid \underline{z}_s]^2\right| &\leq b_s \left\{\mathrm{MDD}[m_s(Y, X, \theta_1) - m_s(Y, X, \theta_2) \mid \underline{z}_s]^2\right\}^{1/2} \\
&\leq b_s c \|\theta_1 - \theta_2\|_{L^2}
\end{aligned}$$

for some finite constants b_s and c, where the first and second inequalities hold by Lemmas A.2 and 3.1, respectively. As a result,

$$\begin{aligned}
|\mathcal{Q}(\theta_1) - \mathcal{Q}(\theta_2)| &= \left|\sum_{s=1}^{T-R} \mathrm{MDD}[m_s(Y, X, \theta_1) \mid \underline{z}_s]^2 - \sum_{s=1}^{T-R} \mathrm{MDD}[m_s(Y, X, \theta_2) \mid \underline{z}_s]^2\right| \\
&\leq \sum_{s=1}^{T-R} \left|\mathrm{MDD}[m_s(Y, X, \theta_1) \mid \underline{z}_s]^2 - \mathrm{MDD}[m_s(Y, X, \theta_2) \mid \underline{z}_s]^2\right| \\
&\leq \left[\sum_{s=1}^{T-R} b_s\right] c \|\theta_1 - \theta_2\|_{L^2}.
\end{aligned}$$

This completes the proof of the lemma. ■

Proof of Lemma A.5. We refer to Conditions (i) – (iv) listed in the statement of Lemma A.5 as C(i) – C(iv), respectively.

By C(iii), $\forall\ \varepsilon > 0$, $\exists$ a constant $L_\varepsilon < \infty$ and a positive integer N_ε s.t.

$$\Pr\left(\sup_{\theta \in \Theta_N} |Q_N(\theta) - Q(\theta)| < L_\varepsilon b_N\right) \geq 1 - \varepsilon \tag{B.14}$$

for all $N \geq N_\varepsilon$. For any given $\theta^0 \in \Theta_I$, it follows from C(ii) that $\exists$ a sequence $\{\theta_N^0\}$ with $\theta_N^0 \in \Theta_N$ s.t. $\mathrm{d}(\theta_N^0, \theta^0) \leq \sigma_N$.

Let $A_\varepsilon \equiv \max\left\{2\sqrt{L_\varepsilon / a_1}, \sqrt{2a_2/a_1}\right\} < \infty$ and $\rho_{N,\varepsilon} \equiv A_\varepsilon \max\{\sigma_N, b_N^{1/2}\}$. Let $\Theta_I^{\rho_{N,\varepsilon}}$ be the open $\rho_{N,\varepsilon}$ enlargement of Θ_I under $\mathrm{d}(\cdot, \cdot)$. By C(i) (compactness of Θ) and C(iv), it holds that

$$\Delta_{N,\varepsilon} \equiv \inf_{\theta \in (\Theta_I^{\rho_{N,\varepsilon}})^c \cap \Theta} Q(\theta) \geq a_1 \rho_{N,\varepsilon}^2,$$

where A^c denotes the complement of set A. Note that

$$\begin{aligned}
Q\left(\hat{\theta}_N\right) - Q\left(\theta_N^0\right) &= \left[Q\left(\hat{\theta}_N\right) - Q_N\left(\hat{\theta}_N\right)\right] + \left[Q_N\left(\hat{\theta}_N\right) - Q_N\left(\theta_N^0\right)\right] + \left[Q_N\left(\theta_N^0\right) - Q\left(\theta_N^0\right)\right] \\
&\leq \left[Q\left(\hat{\theta}_N\right) - Q_N\left(\hat{\theta}_N\right)\right] + \left[Q_N\left(\theta_N^0\right) - Q\left(\theta_N^0\right)\right] \\
&\leq \left|Q\left(\hat{\theta}_N\right) - Q_N\left(\hat{\theta}_N\right)\right| + \left|Q_N\left(\theta_N^0\right) - Q\left(\theta_N^0\right)\right|, \qquad \text{(B.15)}
\end{aligned}$$

where the first inequality holds because $Q_N\left(\hat{\theta}_N\right) - Q_N\left(\theta_N^0\right) \leq 0$ by the definition of $\hat{\theta}_N$. Then for $N \geq N_\varepsilon$, we have

$$\begin{aligned}
\Pr\left(Q\left(\hat{\theta}_N\right) < Q\left(\theta_N^0\right) + \frac{\Delta_{N,\varepsilon}}{2}\right) &= \Pr\left(Q\left(\hat{\theta}_N\right) - Q\left(\theta_N^0\right) < \frac{\Delta_{N,\varepsilon}}{2}\right) \\
&\geq \Pr\left(Q\left(\hat{\theta}_N\right) - Q\left(\theta_N^0\right) < \frac{a_1\rho_{N,\varepsilon}^2}{2}\right) \\
&\geq \Pr\left(\left|Q\left(\hat{\theta}_N\right) - Q_N\left(\hat{\theta}_N\right)\right| + \left|Q_N\left(\theta_N^0\right) - Q\left(\theta_N^0\right)\right| < 2L_\varepsilon b_N\right) \\
&\geq 1 - \varepsilon \qquad \text{(B.16)}
\end{aligned}$$

where the second inequality follows from (B.15) and the fact that $a_1\rho_{N,\varepsilon}^2/2 \geq 2L_\varepsilon b_N$, and the last inequality follows from (B.14).

It follows from C(iv) that $Q\left(\theta_N^0\right) \leq a_2\,\mathrm{d}\left(\theta_N^0, \Theta_I\right)^2 \leq a_2\,\mathrm{d}\left(\theta_N^0, \theta^0\right)^2 \leq a_2\sigma_N^2 \leq \Delta_{N,\varepsilon}/2$ for N large enough, which, together with (B.16), implies that

$$\Pr\left(Q\left(\hat{\theta}_N\right) < \Delta_{N,\varepsilon}\right) \geq \Pr\left(Q\left(\hat{\theta}_N\right) < Q\left(\theta_N^0\right) + \frac{\Delta_{N,\varepsilon}}{2}\right) \geq 1 - \varepsilon \qquad \text{(B.17)}$$

for N large enough.

Note that $Q(\hat{\theta}_N) < \Delta_{N,\varepsilon}$ if and only if $\hat{\theta}_N \in \Theta_I^{\rho_{N,\varepsilon}}$, or, equivalently, $\mathrm{d}(\hat{\theta}_N, \Theta_I) < \rho_{N,\varepsilon} = A_\varepsilon \max\{\sigma_N, b_N^{1/2}\}$. Therefore, we can rewrite (B.17) as

$$\Pr\left(\mathrm{d}\left(\hat{\theta}_N, \Theta_I\right) < A_\varepsilon \max\{\sigma_N, b_N^{1/2}\}\right) \geq 1 - \varepsilon$$

for N large enough. This exactly shows that $\mathrm{d}\left(\hat{\theta}_N, \Theta_I\right) = O_p\left(\max\{\sigma_N, b_N^{1/2}\}\right)$. ■

To prove Lemma A.6, we need the following Lemma.

Lemma B.1 *Consider a generic econometric model $Q(\theta) = 0$, the identified set of which is characterized by $\Theta_I \equiv \{\theta \in \Theta : Q(\theta) = 0 \text{ a.s.}\}$. Suppose the following conditions hold: (i) $Q(\cdot) \geq 0$ and Θ is compact under (pseudo-)norm $\|\cdot\|$; (ii) $\Theta_N \subseteq \Theta$ are closed and s.t. $\exists\ \Pi_N\theta \in \Theta_N$ for each $\theta \in \Theta$ s.t. $\sup_{\theta\in\Theta}\|\Pi_N\theta - \theta\| = o(1)$; (iii) $\sup_{\theta\in\Theta_N}|Q_N(\theta) - Q(\theta)| = o_p(1)$; (iv) $Q(\cdot)$ is continuous w.r.t. $\|\cdot\|$ in Θ. Then for $\hat{\theta}_N \in \operatorname{argmin}_{\theta\in\Theta_N} Q_N(\theta)$, it holds that $\mathrm{d}_{\|\cdot\|}(\hat{\theta}_N, \Theta_I) = o_p(1)$.*

Proof of Lemma B.1. Lemma B.1 is essentially the same as Lemma A.5 in Santos (2012), except that we do not assume the continuity of Q_N w.r.t. $\|\cdot\|$ in Θ_N. A close inspection on the proof of Lemma A.5 in Santos (2012) shows that the continuity of Q_N does not play a role in the proof. In other words, the proof of Lemma A.5 in Santos (2012) works without assuming the continuity of Q_N, and therefore applies directly to proving Lemma B.1 here. ■

Proof of Lemma A.6. We prove parts (i) and (ii) of the Lemma in turn.

Part I. Proof of part (i).

Let $Q(\theta) \equiv \sum_{s=1}^{T-R} \text{MDD}\left[m_s(Y,X,\theta)\,|z_s\right]^2$ and $Q_N(\theta) \equiv \frac{1}{N} S_N(\theta) = \sum_{s=1}^{T-R} \frac{1}{N} S_{Ns}(\theta)$, as in the proof of Theorem 3.2. Our goal is to show that, over the restricted parameter space $\Theta \cap \Theta_R$ under $\|\cdot\|_{L^2}$, $Q(\cdot)$ and $Q_N(\cdot)$ as specified above satisfy Conditions (i)–(iv) in Lemma B.1.

Due to the nonnegativity of MDD, $Q(\cdot) \geq 0$. By Lemma A.3, Θ is compact under $\|\cdot\|_c$ and hence is compact under $\|\cdot\|_{L^2}$, which is weaker than $\|\cdot\|_c$. Since Θ_R is closed due to the continuity of $L(\cdot)$ under Assumption 2.2, $\Theta \cap \Theta_R$ is also compact under $\|\cdot\|_{L^2}$. So Condition (i) in Lemma B.1 is satisfied. Assumption 3.3(i) guarantees Condition (ii) in Lemma B.1 to hold; Condition (iii) in Lemma B.1 holds according to Theorem 3.1. Condition (iv) in Lemma B.1 holds according to Lemma A.4. Consequently, the conclusion in part (i) follows from Lemma B.1.

Part II. Proof of part (ii).

This part of the proof is similar to the proof of Lemma A.3 in Hong (2017) and it goes as follows. For any given $\theta^0 \in \Theta_I$,

$$\begin{aligned}
\left\|\hat{\theta}_N - \theta^0\right\|_{L^2} &\leq \left\|\hat{\theta}_N - \Pi_N\theta^0\right\|_{L^2} + \left\|\Pi_N\theta^0 - \theta^0\right\|_{L^2} \\
&\leq \varrho_N \,\mathrm{d}_w\left(\hat{\theta}_N, \Pi_N\theta^0\right) + \delta_{s,N} \\
&\leq 2\varrho_N\left[\mathrm{d}_w\left(\hat{\theta}_N, \theta^0\right) + \mathrm{d}_w\left(\Pi_N\theta^0, \theta^0\right)\right] + \delta_{s,N} \\
&\leq 2\varrho_N\left[\mathrm{d}_w\left(\hat{\theta}_N, \theta^0\right) + \delta_{w,N}\right] + \delta_{s,N},
\end{aligned}$$

where the first inequality follows from the triangle inequality for $\|\cdot\|_{L^2}$, the second one holds by Definition 3.3 and Assumption 3.3(i), the third one follows from Lemma 3.1, and the last inequality holds by Assumption 3.3(ii). Taking infimum over $\theta^0 \in \Theta_I \cap \Theta_R$ yields

$$\begin{aligned}
\mathrm{d}_{\|\cdot\|_{L^2}}\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) &\leq 2\varrho_N\left[\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) + \delta_{w,N}\right] + \delta_{s,N} \\
&= O_p\left(\varrho_N \,\mathrm{d}_w\left(\hat{\theta}_N, \Theta_I \cap \Theta_R\right) + \delta_{s,N}\right)
\end{aligned}$$

where the equality holds by the fact that $\delta_{w,N} = o\left(N^{-1/2}\right)$. ■

C A discussion on Assumption 3.4

Recall that we have already required in Assumption 3.2 that the eigenvalues of $\mathbb{E}\left[p^{k_N}(x_t)\,p^{k_N'}(x_t)\right]$ for $t = 1, ..., T$ are uniformly bounded and uniformly bounded away from zero. Here, we mainly focus on discussing sufficient conditions for ϱ_N to satisfy Assumption 3.4 for point-identification cases. Under point-identification,

$$\varrho_N = \sup_{\theta \in \Theta_{oN}: \theta \neq \Pi_N\theta^0} \frac{\mathrm{d}_{\|\cdot\|_{L^2}}(\theta, \Pi_N\Theta_I)}{\mathrm{d}_w(\theta, \Pi_N\Theta_I)},$$

with $\Theta_{oN} = \left\{\theta \in \Theta_N : \left\|\theta - \theta^0\right\|_{L^2} \leq \varsigma_N\right\}$ (where $\varsigma_N \downarrow 0$) being the $o(1)$ neighborhood of θ^0 under the L^2 norm. Since a proper completeness condition is necessary for point-identification, we maintain such a condition for most part of the discussion, stated as follows:

> **(Completeness condition)** *For any measurable function* $v(\cdot) : \mathcal{X} \to \mathbb{R}$, $\mathbb{E}\left[v(x_s)\,|\underline{z}_s\right] = 0$ *iff* $v(\cdot) = 0$ *a.s. for* $s = 1, ..., T - R$.

Note that, for any $\theta \in \Theta_N$, we can write $\theta - \Pi_N\theta^0 = \left(\left(\phi - \phi^0\right)', \Delta_N' p^{k_N}\right)'$ for some $\Delta_N \in \mathbb{R}^{k_N}$. By Assumption 3.2, we have

$$\left\|\theta - \Pi_N\theta^0\right\|_{L^2}^2 \asymp \left|\phi - \phi^0\right|^2 + \left|\Delta_N\right|^2. \tag{C.1}$$

C.1 The case of $R = 0$

We consider the special case of $R = 0$ (i.e., no IFEs) and $T = 1$. Generalization to the $R = 0$ and $T > 1$ case is straightforward. Note that, when $R = 0$, $\theta = g(\cdot)$ and $\Theta = \mathcal{G}$. Consequently, $\forall\ \theta \in \Theta_N$ s.t. $\theta \neq \Pi_N\theta^0$, it holds that $\mathrm{d}_w\left(\theta, \Pi_N\theta^0\right)^2 = \mathrm{MDD}\left[\Delta_N' p^{k_N}(x_{i1})\,|z_{i1}\right]^2$ for some $\Delta_N \neq 0$.

For a fixed N, the following three conditions are equivalent: (i) $\mathrm{MDD}\left[\Delta_N' p^{k_N}(x_{i1})\,|z_{i1}\right]^2 > 0$ for all $\Delta_N \neq 0$; (ii) $\mathbb{E}\left[\Delta_N' p^{k_N}(x_{i1})\,|z_{i1}\right] \neq 0$ for all $\Delta_N \neq 0$; (iii) There is no multicollinearity among the elements of the $p^{k_N}(\cdot)$ vector. Note that the equivalence between (ii) and (iii) follows from the completeness condition. The condition of the eigenvalues of $\mathbb{E}\left[p^{k_N}(x_t)\,p^{k_N'}(x_t)\right]$ being bounded away from zero uniformly over N by Assumption 3.2 guarantees that there is no multicollinearity in $p^{k_N}(\cdot)$ even as $N \to \infty$ and $k_N \to \infty$. Consequently, it can be shown that $\mathrm{d}_w\left(\theta, \Pi_N\theta^0\right) \asymp \left\|\theta - \Pi_N\theta^0\right\|_{L^2}$ for $\theta \in \Theta_{oN}$. Therefore, when $R = 1$, Assumption 3.4 holds trivially with $\varrho_N = O(1)$ under Assumption 3.2.

C.2 The case of $R \geq 1$

Now we consider the case of $R = 1$ and $T = 2$. Generalization to the $R \geq 1$ and $T = R + 1$ case is straightforward at the cost of more tedious algebra.

In the $o(1)$ neighborhood of θ^0 under the L^2 norm, which is also the $o(1)$ neighborhood of $\Pi_N\theta^0$ under the L^2 norm by Assumption 3.3(i) ($\left|\Pi_N\theta^0 - \theta^0\right| \downarrow 0$ sufficiently fast),

$$\begin{aligned}
\mathrm{d}_w\left(\theta, \Pi_N\theta^0\right)^2 &= \mathrm{MDD}\left[m_1(Y, X, \theta) - m_1\left(Y, X, \Pi_N\theta^0\right)|\underline{z}_s\right]^2 \\
&= \rho_{11}\left(\theta^0, \theta - \Pi_N\theta^0\right) + o(1),
\end{aligned}$$

where

$$\begin{aligned}
&\rho_{11}\left(\theta^0, \theta - \Pi_N\theta^0\right) \\
\equiv\ & -\mathbb{E}\left[\left[\frac{\partial m_1\left(Y, X, \theta^0\right)}{\partial\theta}\left[\theta - \Pi_N\theta^0\right]\right]\left[\frac{\partial m_1\left(Y^\dagger, X^\dagger, \theta^0\right)}{\partial\theta}\left[\theta - \Pi_N\theta^0\right]\right]\left|z_1 - z_1^\dagger\right|\right] \\
& +2\mathbb{E}\left[\left[\frac{\partial m_1\left(Y, X, \theta^0\right)}{\partial\theta}\left[\theta - \Pi_N\theta^0\right]\right]\left|z_1 - z_1^\dagger\right|\right]\mathbb{E}\left[\left[\frac{\partial m_1\left(Y^\dagger, X^\dagger, \theta^0\right)}{\partial\theta}\left[\theta - \Pi_N\theta^0\right]\right]\right].
\end{aligned}$$

Recall $\frac{\partial m_1(Y,X,\theta^0)}{\partial\theta}\left[\theta-\Pi_N\theta^0\right]=\left(\phi_1-\phi_1^0\right)\left[y_2-g^0\left(x_2\right)\right]-\left[g\left(x_1\right)-\Pi_N g^0\left(x_1\right)\right]-\phi_1^0\left[g\left(x_2\right)-\Pi_N g^0\left(x_2\right)\right]$. It then follows that

$$\rho_{11}\left(\theta^0,\theta-\Pi_N\theta^0\right)=\left(\phi_1-\phi_1^0,\Delta_N'\right)M_N\begin{pmatrix}\phi_1-\phi_1^0\\ \Delta_N\end{pmatrix},$$

where

$$M_N\equiv-\mathbb{E}\left[W_N\left(X,X^\dagger,Y,Y^\dagger\right)\left|z_1-z_1^\dagger\right|\right]+2\mathbb{E}\left[W_N\left(X,X^\dagger,Y,Y^\dagger\right)\right]\mathbb{E}\left[\left|z_1-z_1^\dagger\right|\right]$$

with

$$W_N\left(X,X^\dagger,Y,Y^\dagger\right)\equiv\begin{pmatrix}y_2-g^0\left(x_2\right)\\ -p^{k_N}\left(x_1\right)-\phi_1^0p^{k_N}\left(x_2\right)\end{pmatrix}\begin{pmatrix}y_2^\dagger-g^0\left(x_2^\dagger\right)\\ -p^{k_N}\left(x_1^\dagger\right)-\phi_1^0p^{k_N}\left(x_2^\dagger\right)\end{pmatrix}'.$$

By the definition of the martingale difference divergence matrix (MDDM) in Lee and Shao (2018) (see their Definition 1 and Lemma 1),[14] we can rewrite M_N above as

$$M_N=\mathrm{MDDM}_o\left[\begin{pmatrix}y_2-g^0\left(x_2\right)\\ -p^{k_N}\left(x_1\right)-\phi_1^0p^{k_N}\left(x_2\right)\end{pmatrix}|z_1\right]+M_{N1}$$

with

$$M_{N1}\equiv\mathbb{E}\left[\begin{pmatrix}y_2-g^0\left(x_2\right)\\ -p^{k_N}\left(x_1\right)-\phi_1^0p^{k_N}\left(x_2\right)\end{pmatrix}\right]\mathbb{E}\left[\begin{pmatrix}y_2-g^0\left(x_2\right)\\ -p^{k_N}\left(x_1\right)-\phi_1^0p^{k_N}\left(x_2\right)\end{pmatrix}\right]'\mathbb{E}\left[\left|z_1-z_1^\dagger\right|\right].$$

By Lemma 1 and Theorem 1 in Lee and Shao (2018), $\mathrm{MDDM}_o\left[\begin{pmatrix}y_2-g^0\left(x_2\right)\\ -p^{k_N}\left(x_1\right)-\phi_1^0p^{k_N}\left(x_2\right)\end{pmatrix}|z_1\right]$ is positive semi-definite (p.s.d.). This, in conjunction with the p.s.d. of M_{N1}, implies that M_N is also p.s.d.

Note that the MDDM_o in Lee and Shao (2018) is defined to examine conditional mean independence. To examine conditional mean zero, we shall redefine

$$\begin{aligned}\mathrm{MDDM}\left(V|W\right)&\equiv-\mathbb{E}\left(VV^{\dagger\prime}\left|W-W^\dagger\right|\right)+2\mathbb{E}\left(V\right)\mathbb{E}\left(V^\dagger\right)'\mathbb{E}\left|W-W^\dagger\right|\\ &=\mathrm{MDDM}_o\left(V|W\right)+\mathbb{E}\left(V\right)\mathbb{E}\left(V^\dagger\right)'\mathbb{E}\left|W-W^\dagger\right|.\end{aligned}$$

Then it is straightforward to conclude, based on Theorem 1 in Lee and Shao (2018), that $\forall\ V\in\mathbb{R}^p$ and $W\in\mathbb{R}^q$ s.t. $\mathbb{E}\left(|V|^2+|W|^2\right)<\infty$, $\exists\ p-s$ linearly independent combinations of V s.t. they are conditionally mean zero w.r.t. W, iff $\mathrm{rank}\left(\mathrm{MDDM}\left(V|W\right)\right)=s$. Consequently, M_N is strictly positive definite if and only if:

$$\text{no element of }\mathbb{E}\left[\begin{pmatrix}y_2-g^0\left(x_2\right)\\ -p^{k_N}\left(x_1\right)-\phi_1^0p^{k_N}\left(x_2\right)\end{pmatrix}|z_1\right]\text{ equals zero.}$$

If, in addition, we require that the smallest eigenvalue of M_N be bounded away from zero, a condition similar to Assumption 3.2, then

$$\begin{aligned}\mathrm{d}_w\left(\theta,\Pi_N\theta^0\right)^2&=\rho_{11}\left(\theta^0,\theta-\Pi_N\theta^0\right)+o_p\left(1\right)\\ &\succeq\left|\phi-\phi^0\right|^2+\left|\Delta_N\right|^2,\end{aligned}\tag{C.2}$$

[14] Lee and Shao (2018) extend the $\mathrm{MDD}_o\left(V|W\right)$ concept of Shao and Zhang (2014) for a scalar variable V to $\mathrm{MDDM}_o\left(V|W\right)$ to a vector-valuded V. Specifically, for variables V and W, both of which can be vector-valued, s.t. $\mathbb{E}\left(|V|^2+|W|^2\right)<\infty$, Lee and Shao (2018) specify $\mathrm{MDDM}_o\left(V|W\right)=-\mathbb{E}\left[\left(V-\mathbb{E}\left(V\right)\right)\left(V^\dagger-\mathbb{E}\left(V^\dagger\right)\right)'\left|W-W^\dagger\right|\right]$.

which, in conjunction with (C.1), implies that $\mathrm{d}_w\left(\theta, \Pi_N \theta^0\right)^2 \succeq \left\|\theta - \Pi_N \theta^0\right\|_{L^2}^2$. This, together with Lemma 3.1, implies that $\mathrm{d}_w\left(\theta, \Pi_N \theta^0\right)^2 \asymp \left\|\theta - \Pi_N \theta^0\right\|_{L^2}^2$ in an $o(1)$ neighborhood of θ^0 under the L^2 norm. Consequently, Assumption 3.4 is satisfied with $\varrho_N = O(1)$.

REFERENCES

Hong, S., 2017. Inference in semiparametric conditional moment models with partial identification. Journal of Econometrics 196, 156–179.

Lee, C. E., Shao, X., 2018. Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series. Journal of the American Statistical Association 113, 216-229.

Santos, A., 2012. Inference in nonparametric instrumental variables with partial identification. Econometrica 80, 213–275.

Shao, X., Zhang, J., 2014. Martingale divergence correlation and its use in high dimensional variable screening. Journal of the American Statistical Association 109, 1302-1318.

Su, L., Zheng, X., 2017. A Martingale-difference-divergence-based test for specification. Economics Letters 156, 162–167.